Universitat d'Alacant
Universidad de Alicante

Análisis de datos continuos: Modelos de Análisis de la Varianza y de la Covarianza

Andreu Nolasco

Andreu Nolasco Bonmatí es profesor del Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia de la Universidad de Alicante en el que ha venido desarrollando su labor como docente e investigador. Ha impartido docencia en diversas titulaciones de Ciencias de la Salud (Medicina, Enfermería, Nutrición humana y dietética, Óptica, etc.) tanto en estudios de grado como en posgrado (máster y doctorado), en materias y/o asignaturas como Bioestadística, Estadística Avanzada, Demografía y Salud, Metodología de la Investigación, Desigualdades en Salud, Análisis de la mortalidad, etc. Ha venido desarrollando investigación en líneas como: Análisis de la Mortalidad, Geografía Sanitaria, Estadísticas Sanitarias, Encuestas de salud, Demografía y salud, Desigualdades en salud y otras. Su experiencia en la aplicación del método estadístico en el entorno de las Ciencias de la Salud proviene y se refleja en la dirección de numerosos proyectos de investigación, tesis doctorales, publicaciones científicas y el continuo contacto con el contexto sanitario a través del asesoramiento metodológico a diversas instituciones sanitarias (Administración sanitaria, Centros de Salud y Salud Pública, Hospitales y otras).

SUMARIO

Presentación.. 5
Introducción. Conceptos generales... 5
El Análisis de la Varianza. Consideraciones previas. Requerimientos... 9
Análisis de la Varianza de un factor.. 10
Análisis de la Covarianza 29
Análisis de la varianza de medidas repetidas...................................... 42

PRESENTACIÓN

Esta publicación persigue introducir en la utilización de algunos procedimientos y técnicas de análisis estadístico de datos a aquellas personas que requieran por los objetivos de su investigación de la utilización de modelos de análisis multivariante. Los procedimientos descritos en este trabajo guardan relación con aquellas situaciones en las que la variable principal a estudio (o variable respuesta) es de tipo continuo, mientras que el resto de variables implicadas en el análisis son cuantitativas o cualitativas. En esta situación, los métodos epidemiológicos clásicos tienden a abordar el estudio de las interrelaciones entre variables (asociaciones, confusiones, interacciones, etc.) sugiriendo la categorización de las variables cuantitativas, con la consiguiente pérdida de información, y la utilización de medidas o modelos para datos categóricos, para estimar y cuantificar la asociación entre variables. Como se verá, los conceptos de asociación, confusión o interacción tienen pleno sentido con variables cuantitativas y los procedimientos aquí descritos tienen capacidad para su detección y cuantificación. Aunque no es imprescindible, resulta conveniente que el lector se encuentre familiarizado, al menos a nivel básico, con los conceptos de asociación, confusión e interacción entre variables. Es igualmente deseable que el lector disponga de conocimientos básicos sobre regresión lineal.

En la estructura seguida en la presentación de los procedimientos se parte de una introducción a la situación de análisis, se establece el objetivo y los posibles modelos alternativos que pueden producirse en esa situación, se formula las pruebas de hipótesis para identificar la situación y por último se aplica sobre un ejemplo (los resultados han sido obtenidos utilizando para ello el paquete de aplicaciones estadísticas SPSS®).

INTRODUCCIÓN. CONCEPTOS GENERALES

Buena parte de la investigación en el entorno sanitario persigue establecer y caracterizar las relaciones o asociaciones existentes entre un conjunto de variables. Dentro de los métodos estadísticos que permiten verificar este objetivo, suele denominarse análisis multivariante al conjunto de técnicas y procedimientos que estudian conjuntamente tres o más variables. La estadística matemática ha puesto a disposición de los investigadores una multiplicidad de métodos pensados para dar respuesta al objetivo mencionado. Sin embargo, en la práctica, la utilización de unos u otros procedimientos viene orientada por el papel y el tipo de las variables involucradas en el análisis.

Para introducir estas ideas, considere como ejemplo un estudio en el que sobre una muestra de individuos han sido recogidos datos de las siguientes variables:

URICO	= Nivel de ácido úrico (medido en mg/l)
EDAD	= Edad en años
ALCOHOL	= Consumo de alcohol: 1 'bajo' 2 'moderado/alto'
EDADREC	= Edad en tres categorías '≤30 años', '30-40 años', '>40 años

Suponga que el investigador persigue averiguar si el consumo de alcohol es un factor de riesgo con efecto sobre el ácido úrico pero teniendo en cuenta la edad de los individuos. Podemos establecer varias clasificaciones de esta situación:

- **Clasificación por el papel de las variables**:

Clasificación según el papel en el análisis

- **Variable respuesta:** Aquella sobre la que se produce el efecto de las demás. Sus variaciones son consecuencia de las variaciones de las demás. También es denominada variable dependiente (Ejemplo: ácido úrico)
- **Variables explicativas:** El resto de variables. Se trata de determinar en que forma afectan a la respuesta. También se llaman independientes o predictores (Ejemplo: alcohol, edad)
  - **Factores:** El interés del estudio estriba en detectar caracterizar su relación y efecto con la variable respuesta. (Ejemplo: alcohol)
  - **Covariables:** Son variables incluidas en el análisis por su posible relación con la respuesta y los factores. (Ejemplo: edad)

Figura 1.- Clasificación de variables según su papel en el análisis

Si suponemos que el objetivo del investigador es averiguar si el consumo de alcohol se relaciona con el ácido úrico teniendo en cuenta que la edad es una variable que puede relacionarse con ambas, diremos que el ácido úrico es la variable respuesta, y el consumo de alcohol y la edad son explicativas (pueden influir en el mayor o menor nivel de ácido úrico), aunque de estas dos, el alcohol será el factor (en el lenguaje epidemiológico sería identificada como el *factor de riesgo* a estudiar) y la edad la covariable (en el lenguaje epidemiológico podríamos decir variable de *confusión* o de *interacción*) (Ver figura 1).

Clasificación de las covariables (variables de control)

- **De confusión:** La relación entre el factor y la respuesta es diferente de tener en cuenta en el análisis a la covariable a no hacerlo. Para que una covariable sea de confusión es necesario y suficiente que se relacione con ambas, el factor y la respuesta. Debe tenerse en cuenta que lo que se confunde es la relación entre factor y respuesta
- **De interacción:** La relación entre el factor y la respuesta se modifica según los valores o niveles de la covariable. Diremos que el efecto y/o relación entre factor y variable respuesta es inseparable de la covariable.

Figura 2.- Clasificación de las covariables

Las diferencias en el papel de las variables explicativas se agudizan al considerar el papel de las covariables, también denominadas variables de control (Ver figura 2). En general, su introducción en el análisis obedece a su posible papel como variables de confusión o interacción. La confusión y la interacción son conceptos diferentes pero ambos tienen que ver con el estudio de la relación o efecto entre factores y variables respuesta.

Con independencia de la medida elegida para cuantificar la magnitud de la relación entre estas variables, diremos que existe confusión cuando la interpretación de la relación entre factor y variable respuesta difiere de incluir la covariable en el análisis de los datos a no hacerlo, por ejemplo, la relación entre alcohol y ácido úrico es diferente cuando consideramos el efecto de la edad a cuando no tenemos en cuenta la edad de los sujetos. Por otra parte, diremos que existe interacción cuando la relación entre factor y variable respuesta cambia (se modifica) según los niveles o valores de la covariable, por ejemplo, la relación entre alcohol y ácido úrico cambia (es por ejemplo más intensa) de unas a otras edades (mayor relación según aumenta la edad, por ejemplo).

Debe tenerse en cuenta que la interacción es un efecto jerárquicamente superior a la confusión, es decir, si existe interacción no tiene sentido preguntarse sobre la confusión.

- **Clasificación por el tipo de variables:**

Clasificación según el tipo

- **Variables cuantitativas:** Se presentan en forma de valores numéricos fruto de procesos de medida o recuento (Ejemplo: edad, ácido úrico)
  - **Continuas:** Son puramente resultados de procesos de medida, entre cualesquiera dos valores siempre cabe otro
  - **Discretas:** Surgen de procesos de recuento. Tienen un número finito o infinito numerable de valores
- **Variable cualitativa:** Se presenta en forma de modalidades o estados. (Ejemplo: alcohol). En ocasiones puede proceder de la categorización de variables cuantitativas (Ejemplo: edad en categorías, '≤ 30 años', '30 - 40 años', '> 40 años')

Figura 3.- Clasificación según el tipo de variables

A lo largo de esta publicación vamos a abordar algunas de las técnicas de análisis para aquellas situaciones en las que dispongamos de una variable respuesta de tipo cuantitativa continua y variables explicativas, factores o covariables, tanto cualitativas como cuantitativas (Ver figura 3). Las técnicas abordadas se enmarcan dentro de la clase de los llamados modelos lineales. El caso más conocido de este tipo de modelos es la regresión lineal que no será desarrollada en esta publicación, dedicada a otros procedimientos multivariantes con variable respuesta cuantitativa.

El cuadro 1 recoge los procedimientos que son objeto de estudio, clasificados según el tipo de variables respuesta y explicativas, así como el objetivo del análisis y algunas preguntas 'típicas' asociadas a este objetivo. Las respuestas a éstas y otras preguntas de interés son objeto de la presente publicación.

Cuadro 1.- Procedimientos de análisis desarrollados en esta publicación

Procedimiento	Tipo de variables: Respuesta	Tipo de variables: Explicativas	Objetivo del análisis. Algunas preguntas de interés
ANÁLISIS DE LA VARIANZA	Cuantitativa continua	Una o más variables cualitativas	Comprobar si las medias de la variable cualitativa difieren según los niveles de una o más variables explicativas cualitativas. Detectar si existe confusión o interacción al evaluar el efecto de los factores sobre la variable respuesta (medido a través de diferencias de medias). Algunas preguntas: • ¿Hay diferencias en la media de ácido úrico según consumo de alcohol? • ¿Hay diferencias en la media de ácido úrico según se pertenezca a uno u otro grupo de edad? • Las diferencias en las medias de ácido úrico según consumo de alcohol ¿Son las mismas en todos los grupos de edad?
ANÁLISIS DE LA COVARIANZA	Cuantitativa continua	Al menos una variable cualitativa (generalmente el factor) y una cuantitativa (generalmente la covariable)	Comprobar si existe asociación entre la variable respuesta y un factor, controlando el posible efecto de confusión o interacción de una covariable. Obtener medidas ajustadas por la covariable de las medias de la variable respuesta en los niveles del factor. Algunas preguntas: • El efecto del consumo de alcohol sobre el nivel medio de ácido úrico, ¿es el mismo a cualquier edad? ¿Es el mismo que si no tenemos en cuenta la edad? • Si la diferencia en los valores medios de ácido úrico según consumo de alcohol es la misma a cualquier edad, ¿cuánto vale? • Si las edades de los bebedores y de los no bebedores no son iguales, ¿cuánto valen las medias de ácido úrico ajustadas por edad (si tuviean la misma edad) de ambos grupos? • Si la diferencia en los valores medios de ácido úrico según consumo de alcohol no es constante y depende de la edad, ¿cuánto vale esta diferencia en sujetos de 43 años?
ANÁLISIS DE LA VARIANZA DE MEDIDAS REPETIDAS	Cuantitativa continua	Una variable cualitativa que representa el número de momentos en que se observa la variable respuesta en una muestra de individuos. Otros factores o covariables	Comprobar si las medias de la variable respuesta difieren entre los momentos estudiados. Controlar el efecto de covariables. Algunas preguntas: • Se observa el ácido úrico de un grupo de individuos en tres momentos: Al inicio de un tratamiento, a los tres meses y a los seis meses. Se quiere saber si hay diferencias en las medias entre el inicio, los tres y los seis meses • Las diferencias anteriores, ¿son iguales en hombres y en mujeres?

EL ANÁLISIS DE LA VARIANZA. CONSIDERACIONES PREVIAS. REQUERIMIENTOS

El análisis de la varianza (ANOVA[1]) es una técnica para averiguar cómo afectan una colección de variables explicativas cualitativas a una variable respuesta continua. La forma en la que va a ser evaluado el efecto es a través de las medias de la variable respuesta, pues se tratará de averiguar si existen diferencias en ellas y como se comportan tales diferencias según los niveles de los factores. El nombre de análisis de la varianza hace referencia a que los estadísticos de contrastes de hipótesis construidos para esta técnica utilizarán estimaciones de la varianza.

Como se observa en la formulación de la situación de aplicación del ANOVA, existe una similitud con la regresión lineal. En la práctica suele abordarse el análisis a través de un ANOVA cuando la totalidad de las variables explicativas son cualitativas, mientras que se aborda a través de regresión lineal cuando las variables explicativas son todas cuantitativas o mixtas, cualitativas y cuantitativas (en este caso el análisis de la covarianza será apropiado, como se describirá más adelante). De hecho, el modelo de ANOVA puede ser representado en forma de ecuaciones tipo regresión lineal.

Otra consideración común a todos los modelos de ANOVA guarda relación con la necesidad de distinguir entre lo que se denominan factores fijos y los factores aleatorios. En el ANOVA de factores fijos los niveles de los factores a estudio son los únicos que nos interesan o los únicos que se pueden producir para establecer las diferencias de medias existentes y los efectos en el modelo. Como ejemplo de ello considere, respecto a las variables definidas en el apartado anterior, estudiar si existe diferencia en las medias de ácido úrico según consumo de alcohol. En este caso, el consumo de alcohol (bajo o moderado/alto) incluye todos los niveles posibles de consumo (cada sujeto está necesariamente en una de estas categorías). Hablaremos de ANOVA de efectos aleatorios cuando los niveles de los factores estudiados son una muestra de los posibles niveles que quisiéramos estudiar. Como ejemplo, suponga que queremos estudiar las diferencias en la edad media de pacientes hospitalizados entre hospitales de la provincia de Alicante. Pero disponemos de datos obtenidos sobre 4 hospitales, seleccionados al azar de entre todos los hospitales de la provincia. En este caso el factor hospital posee unos niveles que no son todos los posibles ni incluye todos los deseados.

Por otra parte, se dice que un diseño de ANOVA es completo si todas las casillas que resultan de los cruces de niveles de los factores involucrados tienen datos. Cuando esto no sucede el diseño se considera incompleto.

La distinción entre el carácter de los factores (fijos o aleatorios) y si se trata de un diseño completo o incompleto es importante puesto que influye sobre las pruebas de hipótesis y elementos inferenciales que dan solución al ANOVA.

En los procedimientos descritos en este texto se considerará siempre que los factores son de tipo fijo y el diseño completo. Para aplicaciones con factores aleatorios o diseños incompletos deben ser consultadas otras referencias bibliográficas.

Por último, es necesario establecer los requerimientos, suposiciones necesarias para poder aceptar con rigor los resultados inferenciales en el ANOVA. Para los modelos de factores fijos y diseño completo, las suposiciones necesarias se enuncian como sigue:

• Independencia. Disponemos de muestras aleatorias, una por cada nivel o cruce de niveles de los factores, de observaciones independientes de la variable respuesta

• Normalidad. La variable respuesta sigue un modelo normal en cada subpoblación definida por cada nivel o cruce de niveles de los factores

[1] Del inglés *Analysis of Variance*

• Homogeneidad de varianzas. La varianza de la variable respuesta es la misma en cada subpoblación definida por cada nivel o cruce de niveles de los factores.

En general, algunos de estos requerimientos pueden ser enunciados en función de los errores o residuos de los modelos, digamos $\{e_i\}_{i=1}^{n}$, donde cada residuo e_i es la diferencia entre el valor observado de la variable respuesta y el estimado por el modelo de ANOVA correspondiente. Así, en términos de los residuos los requerimientos serán:

• Normalidad. Los residuos siguen un modelo normal

• Homogeneidad de varianzas. La varianza de los residuos es constante

• Independencia de los residuos. Los residuos son estadísticamente independientes

Algunos de estos requerimientos pueden ser comprobados de forma más sencilla que otros. Así, la normalidad puede ser comprobada a través de los residuos, gráficamente o con alguna prueba no paramétrica como la de Kolmogorov-Smirnov. En cualquier caso, si los tamaños de muestra son grandes (según algunos autores a partir de 20 observaciones en cada subpoblación), el teorema central del límite ofrece buenas garantías si la desviación de la normalidad no es muy acusada. Respecto a la homogeneidad de varianzas, los paquetes de ordenador suelen incorporar en los procedimientos de ANOVA pruebas de hipótesis que formulan como hipótesis nula la homogeneidad y alterna lo contrario. La independencia es un requerimiento de difícil comprobación, aunque la vigilancia sobre este requerimiento debe producirse especialmente en estudios con observaciones repetidas sobre sujetos o con secuencias temporales de algún tipo en las observaciones. En estudios observacionales es poco probable que las observaciones no sean independientes.

Las soluciones a la violación de los requerimientos son diversas, aunque ninguna de ellas contestará exactamente la pregunta realizada con el ANOVA. Así podemos:

• Analizar los datos a través de pruebas no paramétricas
• Eliminar valores extremos
• Categorizar la variable respuesta y utilizar análisis para datos frecuenciales (tablas de contingencia)
• Transformar los datos (logaritmo, raíz cuadrada, etc...)
• Confiar y apelar a la robustez[2] del ANOVA

ANÁLISIS DE LA VARIANZA DE UN FACTOR

El ANOVA de un factor es el modelo básico de análisis de la varianza. No puede considerarse estrictamente un modelo multivariante, al involucrar únicamente a dos variables, la variable respuesta y el factor explicativo.

Los datos

Sean

Y = variable respuesta, cuantitativa continua
X = variable explicativa, categórica con I categorías

[2] En estadística se dice que un procedimiento es robusto si es poco sensible a desviaciones moderadas de los requerimientos básicos

Se dispone de una muestra de n observaciones de la variable Y, que podemos considerar estructuradas de la siguiente forma:

CATEGORÍA FACTOR X				
1	. . .	j	. . .	I
y_{11}	.	y_{j1}	.	y_{I1}
y_{12}	.	y_{j2}	.	y_{I2}
y_{13}	.	y_{j3}	.	y_{I3}
y_{14}	. . .	y_{j4}	. . .	y_{I4}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_{1n_1}	.	y_{jn_j}	.	y_{kn_k}

donde, en general, y_{jh} representa la observación h-ésima (con h = 1,2,...,n_j) en la columna o categoría j, y n_1, n_2, ..., n_I los tamaños de muestra respectivos en las I categorías de la variable X.

De entre las variables descritas en el apartado de introducción, considere como ejemplo las variables:

Y = URICO = Ácido úrico = Variable respuesta
X = EDADREC = Edad recodificada en 3 categorías = Variable explicativa

Se dispone de una muestra de observaciones de ácido úrico en cada grupo de la variable que podemos observar en la tabla 1.

Tabla 1.- Observaciones de ácido úrico sobre un conjunto de individuos, según categoría de edad

CATEGORÍA FACTOR X = EDADREC		
x < 30	30 ≤ x < 40	40 ≤ x
54,72,33,30,47,83 40,50,52,38,54,38 34,49,41,38,39,44 28,40,75,40,47,37 45,59,47,46,44,43 52,58,45,43,34,45 42,40,47,39,52,46 39,49,49,51,53,51 33,40,47,52,51,46 43,43,60,84,42,54 41,52,68,52,51,38, 47,24,37,41,46,47 43,44	65,46,39,34,48 41,35,35,31,66 39,49,64,49,33 42,37,37,43,40 40,60,38,42,58 40,54,38,55,58 48,42,44,41,59 37,41,56,44,39 42,35,45,66,48 22,62,46,46,38 42,30,27,35	25,67,66,52,48 63,65,54,72,62 52,71,45,72,65 61,46,62,52,68 48,50,46,48,50 63,73,54,49,69 62,63,47,71,43 53,55,45,50,48 63,61,72,47,42 68,45,63,68,48 75,73,65,69,74 35,30,98,93
n_1 = 74	n_2 = 54	n_3 = 59

El objetivo. Modelos alternativos

Si suponemos que la variable respuesta tiene medias $\{\mu_i\}_{i=1}^{3}$, se trata de averiguar en cual de las siguientes situaciones nos encontramos:

i. El factor no tiene efecto sobre la variable respuesta

Ejemplo: No existen diferencias en las medias de ácido úrico entre las categorías de edad. Sólo existe una media común:

Tabla de valores de las medias de ácido úrico

CATEGORÍA FACTOR X = EDADREC			
$x < 30$	$30 \le x < 40$	$40 \le x$	Global
49,3	49,3	49,3	49,3

Un modelo que represente esta situación es:

$$\mu_i = \mu \qquad\qquad y_{ij} = \mu + e_{ij}$$

con μ = media de la variable para el total de los datos
μ_i = media de la variable en la categoría i del factor
e_{ij} = fluctuación aleatoria propia de la observación j-ésima de la categoría i

pudiendo decir que una observación cualquiera se puede obtener sumando o restando una cantidad a la media global.

ii. El factor tiene efecto sobre la variable respuesta

Ejemplo: Existen diferencias en las medias de ácido úrico entre las categorías de edad:

Tabla de valores de las medias de ácido úrico

CATEGORÍA FACTOR X = EDADREC			
$x < 30$	$30 \le x < 40$	$40 \le x$	Global
46,6	44,3	57,2	49,3

Un modelo que represente esta situación es:

$$\mu_i = \mu + \beta_i \qquad\qquad y_{ij} = \mu + \beta_i + e_{ij}$$

con μ = media de la variable para el total de los datos
μ_i = media de la variable en la categoría i del factor
β_i = efecto propio de la categoría i
e_{ij} = fluctuación aleatoria propia de la observación j-ésima de la categoría i

pudiendo decir que la media de una categoría se puede obtener sumando o restando una cantidad a la media global, y una observación cualquiera se puede obtener sumando o restando una cantidad a la media de su categoría.

Descomposición de la variabilidad

El análisis de la varianza resolverá la identificación de la situación a través de la descomposición de la variabilidad de la variable respuesta en componentes que permitirán construir las pruebas de hipótesis adecuadas. Considere los siguientes estadísticos muestrales obtenidos a partir de los datos:

$$\overline{Y}_{++} = \frac{\sum_{ij} y_{ij}}{n}$$ media global de la variable respuesta

$$\overline{Y}_{i+} = \frac{\sum_{j} y_{ij}}{n_i}$$ media de la variable respuesta en la categoría i del factor, i=1,2,..., I

Puede demostrarse sin dificultad que la variabilidad de la variable Y puede descomponerse de la siguiente forma

$$SS_{TOTAL} = SS_{\beta} + SS_{e} \quad (1)$$

donde

$SS_{TOTAL} = \sum \left(y_{ij} - \overline{Y}_{++}\right)^2$ Variabilidad total de la variable respuesta

$SS_{\beta} = \sum n_i \left(\overline{Y}_{i+} - \overline{Y}_{++}\right)^2$ Variabilidad explicada por el factor. Puede ser considerada una medida de la variabilidad 'entre o inter ' categorías del factor

$SS_{e} = \sum\sum (y_{ij} - \overline{Y}_{i+})^2$ Variabilidad no explicada por el factor, residual o del error. Puede ser considerada una medida de la variabilidad 'dentro o intra' categorías del factor

De la ecuación (1) se desprende que si SS_{β} es grande comparado con SS_{e}, una parte importante de la variabilidad total será atribuible a las diferencias entre categorías del factor en lugar de ser atribuible a diferencias dentro de las categorías del factor.

Pruebas de hipótesis

De acuerdo con el objetivo del análisis en el ANOVA, dispondremos de dos tipos de pruebas de hipótesis:

• Contraste de hipótesis de igualdad de medias de la variable respuesta en todas categorías del factor, formulando las hipótesis:

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_I \qquad\qquad H_a: \quad \exists \mu_i \quad \text{que difiere de las demás}$$

es decir, igualdad entre todas las medias frente a que al menos una difiera de las demás. Esta prueba de hipótesis se resolverá través de las componentes de variabilidad promediadas por sus correspondientes grados de libertad, comparando la variabilidad explicada frente a la variabilidad residual o del error, produciendo la siguiente tabla del ANOVA:

Fuente de variación	Grados de libertad*	Variabilidad absoluta	Cuadrado medio*	Cociente F	Grados de libertad de F
Efecto factor o Variabilidad inter-grupos	I-1	SS_β	$MS_\beta = SS_\beta / I-1$	$\frac{MS_\beta}{MS_e}$	(I-1,n-I)
Error, Residual o variabilidad intra-grupos	n-I	SS_e	$MS_e = SS_e / n-I$	______	_______
Total	n-1	SS_{TOTAL}	$MS_{TOTAL} = SS_{TOTAL} / n-1$	______	______

(*) Los grados de libertad representan el número de fuentes independientes de variación para esa medida de variabilidad. Los cuadrados medios o medias cuadráticas representan las variabilidades promediadas por los grados de libertad que las producen

El estadístico para el contraste formulado es:

$$F = \frac{MS_\beta}{MS_e}$$

cuya distribución de probabilidad es una F de Snedecor con (I-1, n-I) grados de libertad .

• Contrastes para averiguar cuales son las categorías del factor para las que la media de la variable respuesta difiere de las demás. Estos contrastes sólo tienen sentido si el primer contraste de homogeneidad de medias ha resultado significativo. En ese caso, estamos hablando de realizar todas las comparaciones entre parejas de medias (comparaciones múltiples), es decir, realizar un número de comparaciones igual al de combinaciones del número de categorías del factor tomado de dos en dos, C_k^2 , y formulando para cada uno de ellos las hipótesis:

$$H_0: \quad \mu_i = \mu_j \qquad\qquad H_a: \quad \mu_i \neq \mu_j$$

Tales contrastes no deben ser resueltos a través de la prueba t para comparación de medias. El motivo reside en la modificación del nivel de significación real por el hecho de resolver múltiples contrastes, aumentando la probabilidad de encontrar diferencias debidas al azar (mayor probabilidad de error tipo I de la establecida en cada contraste). Una forma sencilla de abordar estas comparaciones es a través del llamado método de Bonferroni. Su aplicación consiste en modificar el nivel de significación de cada contraste individual. Si partimos de un nivel de significación real deseado digamos α, puede ser demostrado que en h comparaciones el máximo posible valor para la significación global es $h\alpha$. Así, puede ser corregido el nivel de significación de cada contraste individual como:

$$\alpha' = \frac{\alpha}{C_k^2}$$

Pero el método de Bonferroni tiene el inconveniente de que el verdadero nivel de significación podría ser mucho menor que el máximo valor utilizado, dado que se trata de una aproximación. Han sido desarrollados diferentes procedimientos que mejoran la aproximación de Bonferroni para producir comparaciones *honestas* entre las medias de las categorías de las variables. En el cuadro 2 se presenta dos de los procedimientos más utilizados, resumiendo sus características de utilización. Tales métodos pueden ser utilizados tanto en contrastes de hipótesis como para construir intervalos de confianza para la diferencia de medias.

Cuadro 2.- Métodos más utilizados en comparaciones múltiples de medias

PROCEDIMIENTO	CONDICIONES DE APLICACION
Método de Tukey	Recomendable si los tamaños de las muestras de cada nivel del factor son iguales, $n_i = n^*$, i=1,2,..., k, y las comparaciones deseadas son entre parejas de medias (no otras funciones de las medias)
Método de Scheffé	Recomendable si los tamaños de las muestras de cada nivel difieren o deseamos comparaciones entre combinaciones funciones de las medias (p.ej. En un caso de 4 categorías del factor, contrastar si el promedio de las dos primeras medias es igual, o diferente, del promedio del resto)

Un ejemplo

Considere como ejemplo la situación y los datos recogidos en la tabla 1. En ella se clasifican 187 observaciones de ácido úrico según las categorías de edad de los sujetos. Aunque la edad es una variable continua, se ha categorizado para ejemplificar situaciones analizables a través de ANOVA.

La tabla siguiente muestra los resultados descriptivos obtenidos para la variable URICO en cada categoría de edad y en el conjunto total de los datos:

URICO

					Intervalo de confianza para la media al 95%			
	N	Media	Desviación típica	Error típico	Límite inferior	Límite superior	Mínimo	Máximo
Edad<30	74	46.59	10.794	1.255	44.09	49.10	24	84
30<=Edad<40	54	44.28	10.267	1.397	41.48	47.08	22	66
40<=Edad	59	58.37	13.538	1.762	54.84	61.90	25	98
Total	187	49.64	13.006	.951	47.77	51.52	22	98

Puede observarse como los valores medios muestrales de ácido úrico son muy similares en las dos primeras categorías de edad, presentando diferencias importantes con la última categoría de edad.

• Primer contraste de hipótesis: ¿Podemos afirmar que existen diferencias significativas entre las medias poblacionales de ácido úrico entre las categorías de edad?. Se trata de resolver el contraste de hipótesis con formulación:

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 \qquad\qquad H_a: \quad \exists \mu_i \quad \text{que difiere de las demás}$$

La tabla siguiente presenta los resultados de la descomposición de la variabilidad

URICO

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	6738.527	2	3369.263	25.076	.000
Intra-grupos	24722.468	184	134.361		
Total	31460.995	186			

El valor del estadístico F, cociente de las medias cuadráticas Inter-grupos e Intra-grupos, 25,076, contrastado en una F de Snedecor con 2 y 184 grados de libertad resulta con un valor de p<0,001 (el valor ,000 en la salida sólo permite establecer que el decimal correspondiente se sitúa a partir de la 4ª posición). Este valor nos conducirá al rechazo de la hipótesis nula y el establecimiento de que al menos una de las medias difiere de las demás.

• Segundo contraste de hipótesis: ¿Entre qué categorías de edad existe diferencia de medias de ácido úrico?. Se trata de resolver las $C_3^2 = 3$ comparaciones de dos medias, a saber:

$$H_0: \quad \mu_1 = \mu_2 \qquad H_a: \quad \mu_1 \neq \mu_2$$
$$H_0: \quad \mu_1 = \mu_3 \qquad H_a: \quad \mu_1 \neq \mu_3$$
$$H_0: \quad \mu_2 = \mu_3 \qquad H_a: \quad \mu_2 \neq \mu_3$$

La siguiente tabla presenta los resultados obtenidos al aplicar los procedimientos de Tukey, Scheffé y Bonferroni para las comparaciones múltiples. En esta situación, según el cuadro 2, la prueba de Scheffé es preferible a Tukey debido a que los tamaños muestrales no son iguales:

Comparaciones múltiples

Variable dependiente: URICO

	(I) EDADREC	(J) EDADREC	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior
DHS de Tukey	Edad<30	30<=Edad<40	2.32	2.075	.505	-2.59	7.22
		40<=Edad	-11.78	2.023	.000	-16.56	-7.00
	30<=Edad<40	Edad<30	-2.32	2.075	.505	-7.22	2.59
		40<=Edad	-14.10	2.183	.000	-19.25	-8.94
	40<=Edad	Edad<30	11.78	2.023	.000	7.00	16.56
		30<=Edad<40	14.10	2.183	.000	8.94	19.25
Scheffe	Edad<30	30<=Edad<40	2.32	2.075	.537	-2.80	7.44
		40<=Edad	-11.78	2.023	.000	-16.77	-6.79
	30<=Edad<40	Edad<30	-2.32	2.075	.537	-7.44	2.80
		40<=Edad	-14.10	2.183	.000	-19.48	-8.71
	40<=Edad	Edad<30	11.78	2.023	.000	6.79	16.77
		30<=Edad<40	14.10	2.183	.000	8.71	19.48
Bonferroni	Edad<30	30<=Edad<40	2.32	2.075	.797	-2.70	7.33
		40<=Edad	-11.78	2.023	.000	-16.67	-6.89
	30<=Edad<40	Edad<30	-2.32	2.075	.797	-7.33	2.70
		40<=Edad	-14.10	2.183	.000	-19.37	-8.82
	40<=Edad	Edad<30	11.78	2.023	.000	6.89	16.67
		30<=Edad<40	14.10	2.183	.000	8.82	19.37

Como se observa en la tabla, las tres pruebas (Tukey, Scheffé y Bonferroni) coinciden en los resultados, aunque no en los valores de p ni en los intervalos de confianza (que son ajustados por comparaciones múltiples), encontrando los siguientes subconjuntos homogéneos (para los que se acepta la igualdad de medias) con nivel de significación 0,05:

URICO

	EDADREC	N	Subconjunto 1	2
DHS de Tukey	30<=Edad<40	54	44.28	
	Edad<30	74	46.59	
	40<=Edad	59		58.37
	Significación		.512	1.000
Scheffe	30<=Edad<40	54	44.28	
	Edad<30	74	46.59	
	40<=Edad	59		58.37
	Significación		.544	1.000

Así, existen diferencias significativas en las medias de ácido úrico entre la categoría de edad de mayores o iguales a 40 años y cualquiera de las demás, pero no entre las dos primeras.

Respecto a los requerimientos, la tabla siguiente muestra el resultado de la prueba de homogeneidad de varianzas:

URICO

Estadístico de Levene	gl1	gl2	Sig.
13,742	2	184	,000

a partir de la cual debemos pensar que se viola la suposición de homogeneidad de varianzas (p<0,001). La normalidad ha sido contrastada a través de los residuos del modelo, con la prueba de Kolmogorov-Smirnov, obteniendo:

Prueba de Kolmogorov-Smirnov para una muestra

		RES_1 Residuo para URICO
N		187
Parámetros normales	Media	.0000
	Desviación típica	11.52894
Diferencias más extremas	Absoluta	.079
	Positiva	.079
	Negativa	-.060
Z de Kolmogorov-Smirnov		1.080
Sig. asintót. (bilateral)		.194

resultado que no permite rechazar la normalidad de la distribución.

A fin de corregir la falta de homogeneidad de varianzas se han realizado dos transformaciones de la variable URICO a través de su logaritmo neperiano y su raíz cuadrada. Se ha replicado el ANOVA con las nuevas variables, LOGURICO y SQURICO, y verificado los requerimientos. Como se observa en las tablas siguientes, los resultados del ANOVA son siginificativos y los requerimientos se verifican para ambas variables.

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
LOGURICO	Inter-grupos	2.488	2	1.244	23.238	.000
	Intra-grupos	9.848	184	.054		
	Total	12.336	186			
SQURICO	Inter-grupos	31.988	2	15.994	24.546	.000
	Intra-grupos	119.890	184	.652		
	Total	151.877	186			

Resultados significativos para las dos transformaciones de la variable URICO

Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
LOGURICO	1.172	2	184	.312
SQURICO	2.539	2	184	.082

No puede rechazarse la homogeneidad de varianzas en ninguna de las transformaciones

Prueba de Kolmogorov-Smirnov para una muestra

		RES_1 Residuo para LOGURICO	RES_2 Residuo para SQURICO
N		187	187
Parámetros normales	Media	.0000	.0000
	Desviación típica	.23010	.80285
Diferencias más extremas	Absoluta	.068	.058
	Positiva	.053	.057
	Negativa	-.068	-.058
Z de Kolmogorov-Smirnov		.935	.791
Sig. asintót. (bilateral)		.347	.559

Y tampoco puede rechazarse la normalidad de los residuos de ninguna de las transformaciones.

El procedimiento utilizado con SPSS® para generar estos resultados ha sido el ANOVA de un factor, incluido en Comparar Medias del desplegable general del menú Analizar. La sintaxis del procedimiento es:

```
ONEWAY
urico BY edadrec
/STATISTICS DESCRIPTIVES HOMOGENEITY
/MISSING ANALYSIS
/POSTHOC = TUKEY SCHEFFE BONFERRONI ALPHA(.05).
```

El procedimiento utilizado para la comprobación de la normalidad de los residuos ha sido la prueba no paramétrica de Kolmogorv-Smirnov con sintaxis:

```
NPAR TESTS  /K-S(NORMAL)= res_1/MISSING ANALYSIS.
```

ANÁLISIS DE LA VARIANZA DE DOS O MÁS FACTORES

Este procedimiento es, en principio, una generalización del ANOVA de un factor. No obstante, como se explicará más adelante, esta generalización no es trivial, y la elección de los mecanismos de descomposición de la variabilidad dependerá fuertemente de los tamaños muestrales de los niveles de los factores y de los cruces entre niveles de factores. Se desarrolla a continuación el caso de dos factores, cuya generalización a tres o más factores resultará por generalización de los conceptos y procedimientos expuestos.

Los datos

Sean

Y = variable respuesta, cuantitativa continua
X_1 = variable explicativa, categórica, con I categorías
X_2 = variable explicativa, categórica, con J categorías

disponiendo de una muestra de n observaciones de la variable Y, distribuidas según la siguiente estructura:

		Factor X_2				
		1	...	j	...	J
	1	y_{11k} (n_{11})				
	. . .					
Factor X_1	i			y_{ijk} (n_{ij})		
	. . .					
	I					y_{IJk} (n_{IJ})

donde, en general, y_{ijk} representa la observación k-ésima (con k = 1,2,..., n_{ij}) en la casilla o cruce de niveles de los factores (i,j), y

n_{ij} = tamaño muestral del cruce de niveles (i,j) de los factores,

$n_{i+} = \sum_j n_{ij}$ = tamaño muestral del nivel i del factor estructurado por filas X_1

$n_{+j} = \sum_i n_{ij}$ = tamaño muestral del nivel j del factor estructurado por columnas X_2

$n_{++} = \sum_i \sum_j n_{ij} = n$ = tamaño muestral total

De entre las variables descritas en el apartado de introducción, considere como ejemplo las variables:

Y = URICO = Ácido úrico = Variable respuesta
X_1 = EDADREC = Edad recodificada en 3 categorías = Variable explicativa
X_2 = ALCOHOL = Bajo, Moderado/Alto = Variable explicativa

Se dispone de una muestra de observaciones de ácido úrico en cada cruce de niveles de los factores. En la tabla 2 podemos observar los tamaños muestrales en cada cruce de factores, por nivel de cada factor y en total.

Tabla 2.- Tamaños muestrales según niveles de los factores

X_1=EDADREC	ALCOHOL 1 = bajo	ALCOHOL 2=moderado/alto	Totales
$x_1 < 30$	$n_{11} = 52$	$n_{12} = 22$	$n_{1+} = 74$
$30 \leq x_1 < 40$	$n_{21} = 42$	$n_{22} = 12$	$n_{2+} = 54$
$40 \leq x_1$	$n_{31} = 29$	$n_{32} = 30$	$n_{3+} = 59$
Totales	$n_{+1} = 123$	$n_{+2} = 64$	$n_{++} = 187$

El efecto del diseño. Los tamaños muestrales en el ANOVA de dos factores

El análisis de la varianza es un procedimiento de análisis con aplicación tanto en diseños experimentales como en diseños observacionales. Sin embargo, el control del investigador al definir el número de sujetos que serán asignados a uno u otros niveles o cruces de niveles cuando el diseño es experimental hace que en estas situaciones su aplicación sea óptima. Esto es así porque la descomposición de la variabilidad en el ANOVA, fundamento de su solución, verificará ecuaciones básicas cuando se den ciertas condiciones entre los tamaños de las muestras y submuestras de niveles y de cruces de niveles, mientras que cuando esto no ocurra, como sucederá en diseños observacionales, habrá que tener la precaución de utilizar las aproximaciones adecuadas para la situación en la que nos encontremos.

En el cuadro 3 podemos observar los tipos de diseños de ANOVA en función de los tamaños muestrales:

Cuadro 3.- Tipos de diseños en ANOVA

DISEÑOS DE ANOVA	CONSIDERACIONES
ANOVA con diseño equilibrado o balanceado: La frecuencia de sujetos en cada casilla es función de los totales marginales $$n_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$	Son casos particulares de diseño equilibrado los de igual número de sujetos por casilla, igual número por columna con número proporcional por fila, e igual número por fila con número proporcional por columna.
ANOVA con diseño desequilibrado o no balanceado: Las frecuencias de las casillas no cumplen la condición de un diseño balanceado. Tienen una distribución desigual que no responde a ningún patrón	Sólo en los diseños balanceados se puede descomponer la variabilidad de forma aditiva. En otro caso, debe corregirse la descomposición de la variabilidad. La mayoría de programas de ordenador distinguen en la forma de descomponer la variabilidad a través de las opciones para modificar la forma de construir las sumas de cuadrados, debiendo atender a si estamos en un diseño balanceado o no.

Hay que tener en cuenta que gran parte de los estudios epidemiológicos o biosanitarios, son de tipo observacional. En estos estudios se produce con facilidad una situación de tamaños muestrales desiguales y no equilibrados. Esto sucede porque los factores no son categorizados hasta haber recolectado los datos, o porque se consideran o se crean nuevas variables después de la obtención de los datos, o porque no sabemos de antemano qué cruces de niveles pueden tener interés. En estas situaciones, el ANOVA de dos factores o multifactorial (tres o más factores) puede ser utilizado pero con la prevención de seleccionar una forma de descomposición de la variabilidad adecuada.

En la exposición que se desarrolla a continuación se supondrá que estamos ante un diseño equilibrado, a efectos de ejemplificación de los modelos y situaciones alternativas y las pruebas para su detección. En la aplicación a un ejemplo se indicará como seleccionar con SPSS® la descomposición adecuada de la variabilidad.

El objetivo. Modelos alternativos

Si suponemos que la variable respuesta tiene medias $\{\mu_{ij}\}_{i=1,j=1}^{i=I,j=J}$ Se pretende averiguar el efecto de las variables explicativas sobre la variable respuesta a través de las diferencias inducidas sobre su media por los factores que clasifican las situaciones

Se trata de averiguar en cual de las siguientes situaciones nos encontramos:

i. Ninguno de los factores tiene efecto sobre la variable respuesta

Ejemplo: No existen diferencias en las medias de ácido úrico ni por categoría de edad, ni por alcohol. Sólo existe una media común

Tabla de valores de las medias de ácido úrico

	ALCOHOL		
X_1=EDADREC	1 = bajo	2=moderado/alto	Totales
$x_1 < 30$	49,3	49,3	49,3
$30 \le x_1 < 40$	49,3	49,3	49,3
$40 \le x_1$	49,3	49,3	49,3
Totales	49,3	49,3	49,3

Un modelo para esta situación es:

$$\mu_{i+} = \mu \qquad \mu_{+j} = \mu \qquad \mu_{ij} = \mu \qquad y_{ijk} = \mu + e_{ijk}$$

con μ = media de la variable para el conjunto de los datos
μ_{i+}= media de la variable en la categoría i del factor X_1
μ_{+j}= media de la variable en la categoría j del factor X_2
μ_{ij} = media de la variable en la combinación (i,j) de categorías de los factores

De esta forma una observación cualquiera se puede obtener sumando o restando una cantidad a la media global

ii. Existen diferencias para uno de los factores, por ejemplo el factor fila X_1

Ejemplo: Hay diferencias en las medias de ácido úrico entre los grupos definidos por consumo de alcohol pero no entre los grupos de edad

Tabla de valores de las medias de ácido úrico

X_1=EDADREC	ALCOHOL 1 = bajo	2=moderado/alto	Totales
$x_1 < 30$	46,6	46,6	46,6
$30 \le x_1 < 40$	44,3	44,3	44,3
$40 \le x_1$	57,2	57,2	57,2
Totales	49,3	49,3	49,3

El modelo para esta situación es:

$$\mu_{ij} = \mu + \alpha_i \qquad\qquad y_{ijk} = \mu + \alpha_i + e_{ijk}$$

con α_i efecto atribuido al nivel i del factor fila

iii. Existen diferencias para ambos factores, siendo las diferencias entre los niveles de un factor las mismas en cualquiera de los niveles del otro factor

Ejemplo: Hay diferencias en las medias de ácido úrico tanto entre los grupos definidos por alcohol como entre los grupos definidos por edad

Tabla de valores de las medias de ácido úrico

X_1=EDADREC	ALCOHOL 1 = bajo	2=moderado/alto	Totales
$x_1 < 30$	36,6	56,6	46,6
$30 \le x_1 < 40$	34,3	54,3	44,3
$40 \le x_1$	47,2	67,2	57,2
Totales	42,5	62,5	49,3

El modelo para esta situación es:

$$\mu_{ij} = \mu + \alpha_i + \beta_j \qquad\qquad y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

con α_i efecto atribuido al nivel i del factor fila
β_i efecto atribuido al nivel j del factor columna

iv. Las diferencias en las medias, o su ausencia, entre los niveles de uno de los factores se modifican según los niveles del otro factor. Este efecto se define como **interacción**

Ejemplo: Existen diferencias en las medias de ácido úrico entre los niveles de alcohol, pero estas diferencias cambian según el grupo de edad en que nos encontremos. Simultáneamente, las diferencias entre los grupos de edad cambiarán según el grupo de alcohol en que nos encontremos

Tabla de valores de las medias de ácido úrico

X_1=EDADREC	ALCOHOL 1 = bajo	2=moderado/alto	Totales
$x_1 < 30$	41,5	58,6	46,6
$30 \le x_1 < 40$	39,7	60,3	44,3
$40 \le x_1$	42,9	70,9	57,2
Totales	41,2	64,7	49,3

El modelo para esta situación es:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \;; \qquad y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

con α_i efecto atribuido al nivel i del factor fila

β_i efecto atribuido al nivel j del factor columna

γ_{ij} efecto atribuible a los niveles i del factor fila, j del factor columna, al ser considerados conjuntamente

La existencia de interacción establece un efecto conjunto de los factores, que interactúan para producir comportamientos específicos y diferenciados según sus niveles conjuntos. La interacción es un efecto de rango superior a la existencia de diferencias de primer orden (para uno u otro de los factores). Si existe interacción lógicamente cada una de las variables tiene efecto que no se puede separar de la otra (no podemos establecer la diferencia de medias de ácido úrico entre las categorías de alcohol sin saber en qué categoría de edad nos encontramos, pues tales diferencias cambian de una a otra categoría de edad)

Descomposición de la variabilidad

El análisis de la varianza resolverá la identificación de la situación a través de la descomposición de la variabilidad de la variable respuesta en componentes que permitirán construir las pruebas de hipótesis adecuadas. Si el diseño de ANOVA es equilibrado, puede demostrarse sin dificultad que la variabilidad de la variable respuesta,Y, puede descomponerse de la siguiente forma

$$SS_{TOTAL} = SS_{\alpha} + SS_{\beta} + SS_{\gamma} + SS_{e} \qquad (2)$$

Donde, de forma semejante al caso del ANOVA de un factor, y definiendo los estadísticos muestrales

$\overline{Y}_{+++} = \dfrac{\sum_{ijk} y_{ijk}}{n_{++}}$ media global de la variable respuesta

$\overline{Y}_{i++} = \dfrac{\sum_{jk} y_{ijk}}{n_{i+}}$ media de la variable respuesta en la categoría i del factor fila, i=1,2,..., I

$\overline{Y}_{+j+} = \dfrac{\sum_{ik} y_{ijk}}{n_{+j}}$ media de la variable respuesta en la categoría j del factor columna, j=1,2,..., J

$\overline{Y}_{ij+} = \dfrac{\sum_{k} y_{ijk}}{n_{ij}}$ media de la variable respuesta en la casilla correspondiente a la categoría i del factor fila y la categoría j del factor columna, i=1,2,..., I; j=1,2,..., J

tendremos las siguientes sumas de cuadrados:

$SS_{TOTAL} =$	$\sum\left(y_{ijk} - \overline{Y}_{+++}\right)^2$	Variabilidad total de la variable Y
$SS_{\alpha} =$	$\sum n_{i+}\left(\overline{Y}_{i++} - \overline{Y}_{+++}\right)^2$	Variabilidad explicada por el factor fila
$SS_{\beta} =$	$\sum n_{+j}\left(\overline{Y}_{+j+} - \overline{Y}_{+++}\right)^2$	Variabilidad explicada por el factor columna
$SS_{\gamma} =$	$\sum\sum n_{ij}\left(\overline{Y}_{ij+} - \overline{Y}_{i++} - \overline{Y}_{+j+} + \overline{Y}_{+++}\right)^2$	Variabilidad explicada por la interacción
$SS_{e} =$	$\sum\sum\sum\left(y_{ijk} - \overline{Y}_{ij+}\right)^2$	Variabilidad no explicada, residual, o de error

A partir de la ecuación (2), una forma de evaluar la magnitud de los efectos del factor fila, del factor columna o de la interacción, será comparar su correspondiente suma de cuadrados con la suma de cuadrados del error.

Pruebas de hipótesis

Los contrastes de hipótesis sobre los efectos existentes se construirán a partir de las componentes de variabilidad promediadas por sus correspondientes grados de libertad (cuadrados medios o medias de cuadrados) aportadas por cada uno de los efectos, en relación a la variabilidad residual o del error, produciendo la siguiente tabla del ANOVA:

Fuente de variación	Grados de libertad	Variabilidad absoluta	Cuadrado medio	Cociente F	Grados de libertad de F
Constante*	1	$SS_\mu = n\bar{Y}^2_{+++}$	MS_μ	MS_μ	(1, n–IJ)
Efecto factor fila	I-1	SS_α	$MS_\alpha = SS_\alpha / (I-1)$	$\frac{MS_\alpha}{MS_e}$	(I-1,n-IJ)
Efecto factor columna	J-1	SS_β	$MS_\beta = SS_\beta / (J-1)$	$\frac{MS_\beta}{MS_e}$	(J-1,n-IJ)
Efecto de interacción	(I-1)(J-1)	SS_γ	$MS_\gamma = SS_\gamma / ((I-1)(J-1))$	$\frac{MS_\gamma}{MS_e}$	((I-1)(J-1), (n-IJ))
Error o Residual	n-IJ	SS_e	$MS_e = SS_e / (n-IJ)$	———	———
Total	n-1	SS_{TOTAL}	$MS_{TOTAL} = SS_{TOTAL} / (n-1)$	———	———

(*) La constante suele ser incluida en las salidas de ordenador. Con ella se comprueba únicamente si la media global es 0

A partir de esta tabla general de ANOVA, se formulan los siguientes contrastes de hipótesis

- Contraste sobre el efecto de la variable fila (F). Se formulan las hipótesis:

H_0 : La variable fila no tiene efecto. Las medias de la variable respuesta no difieren entre sus niveles ($\alpha_i = 0, \ \forall i$)

H_a : Hay diferencias en la variable respuesta entre los niveles de la variable fila ($\exists \alpha_i \neq 0$)

El contraste se resuelve a partir del estadístico calculado como cociente de cuadrados medios:

$F = \frac{MS_\alpha}{MS_e}$, cuya distribución será una F de Snedecor con (I-1, n-IJ) grados de libertad

- Contraste sobre el efecto de la variable columna (C). Se formulan las hipótesis:

H_0 : La variable columna no tiene efecto. Las medias de la variable respuesta no difieren entre sus niveles ($\beta_j = 0, \ \forall j$)

H_α : Hay diferencias en la variable respuesta entre los niveles de la variable fila ($\exists \beta_j \neq 0$)

El contraste se resuelve a partir del estadístico calculado como cociente de cuadrados medios:

$F = \frac{MS_\beta}{MS_e}$, cuya distribución será una F de Snedecor con (J-1, n-IJ) grados de libertad

- Contraste sobre el efecto de interacción (IT). Se formulan las hipótesis:

H_0 : No existe interacción. De ser significativo alguno de los contrastes anteriores, las diferencias son las mismas para cualquiera de los niveles del otro factor ($\gamma_{ij} = 0, \forall i, \forall j$)

H_1 : Existe interacción ($\exists \gamma_{ij} \neq 0$). La significación o no de los contrastes anteriores es irrelevante, pues las diferencias por un factor dependen del nivel del otro factor

El contraste se resuelve a partir del estadístico calculado como cociente de cuadrados medios:

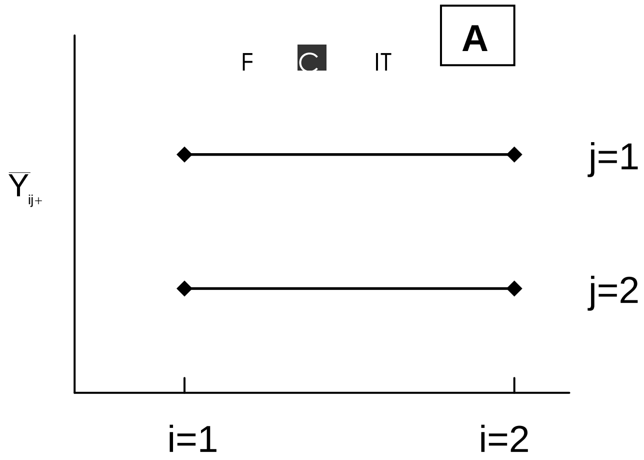
$F = \frac{MS_\gamma}{MS_e}$, cuya distribución será una F de Snedecor con ((I-1)(J-1),(n-IJ)) grados de libertad

Parece lógico resolver en primer lugar el contraste de interacción, puesto que de resultar significativo ya no proceden los contrastes de los efectos fila y columna.

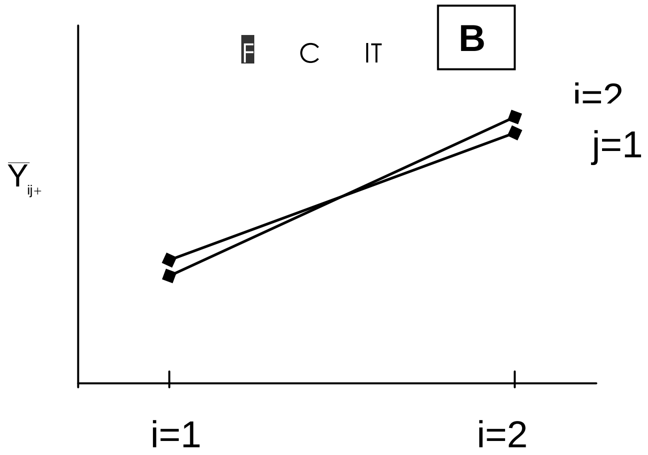
Para identificar las situaciones representadas por los contrastes de hipótesis según su resultado, denominemos por:

F = Efecto fila no significativo	F = Efecto fila significativo
C = Efecto columna no significativo	C = Efecto columna significativo
IT = Interacción no significativa	IT = Interacción significativa

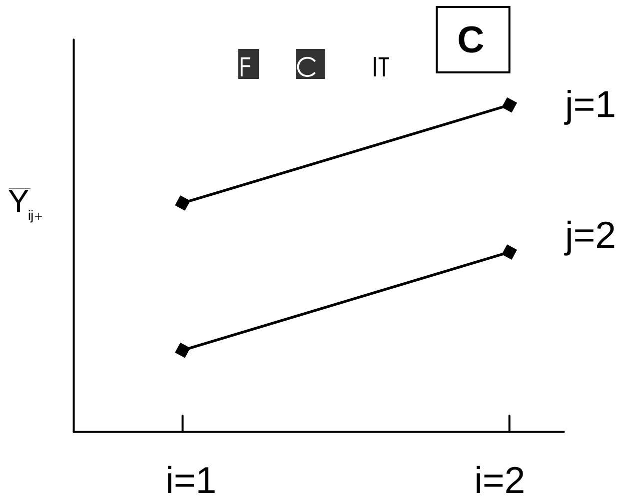
considerando, a modo de ejemplo, una variable s fila y columna de dos categorías. En los gráficos que se presenta a continuación puede observarse como se identifican las diferentes situaciones de los contrastes con los valores de las medias de la variable respuesta según los niveles de los factores. La variable fila se sitúa en el eje de abcisas, con líneas distintas para las categorías de la variable columnas.



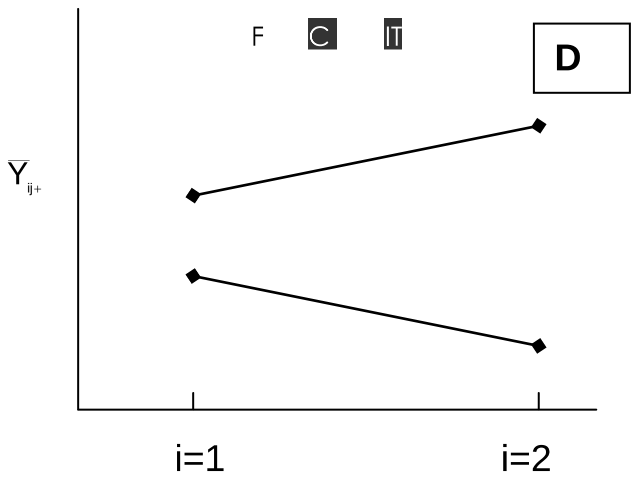
Efecto de la variable columna



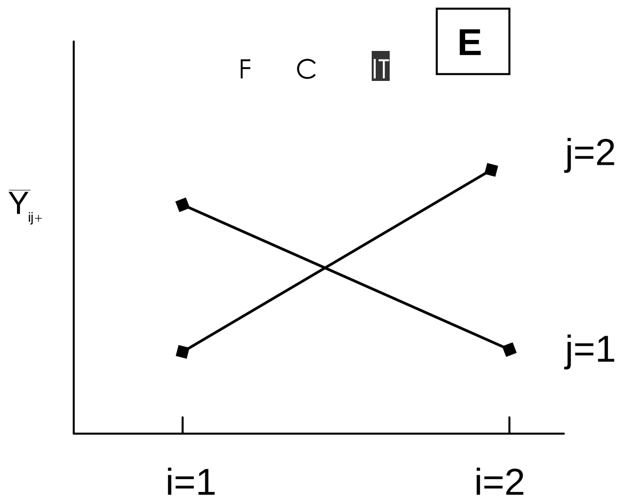
Efecto de la variable fila



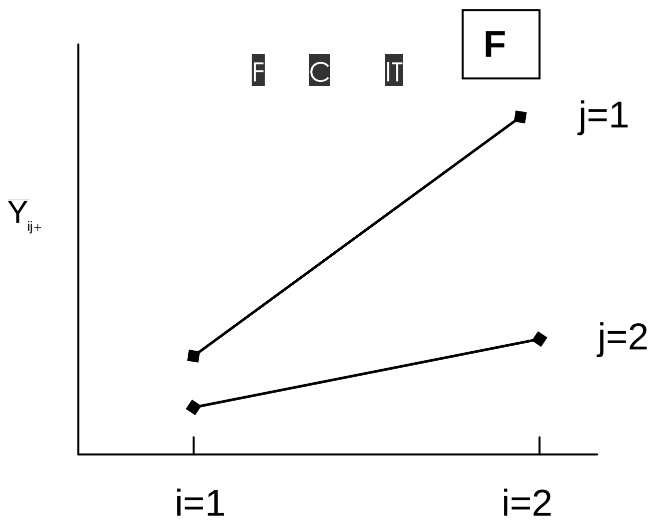
Efecto de ambas variables



Interacción



Interacción



Interacción

Figura 4.- Representación de las medias de la variable respuesta en las situaciones correspondientes a los diferentes efectos de las variables.

Un ejemplo

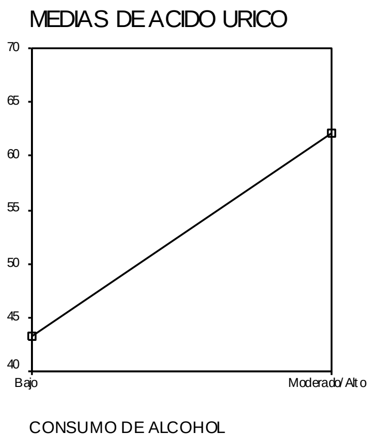
Considere como ejemplo la situación y los datos recogidos en la tabla 1. Se clasifican las 187 observaciones de ácido úrico según las categorías de edad y consumo de alcohol de los sujetos. Los resultados descriptivos se muestran en la tabla siguiente:

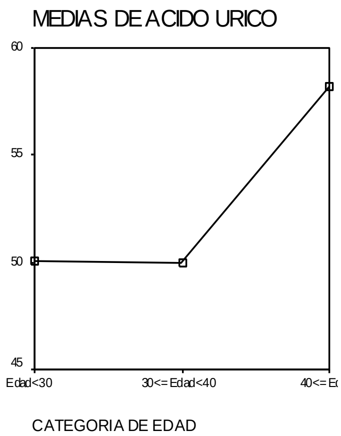
Estadísticos descriptivos

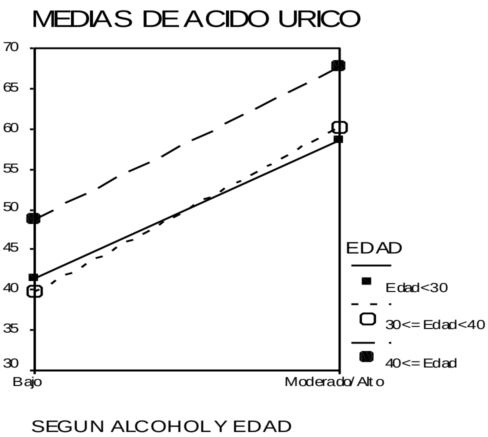
Variable dependiente: URICO

ALCOHOL	EDADREC	Media	Desv. típ.	N
1	Edad<30	41,50	5,599	52
	30<=Edad<40	39,71	5,944	42
	40<=Edad	48,72	11,934	29
	Total	42,59	8,364	123
2	Edad<30	58,64	10,617	22
	30<=Edad<40	60,25	4,288	12
	40<=Edad	67,70	6,778	30
	Total	63,19	8,972	64
Total	Edad<30	46,59	10,794	74
	30<=Edad<40	44,28	10,267	54
	40<=Edad	58,37	13,538	59
	Total	49,64	13,006	187

Las medias observadas sugieren posibles efectos de las variables edad, alcohol o interacción entre ambas. Veamos los gráficos marginales y conjunto:







Los gráficos sugieren posibles efectos de alcohol y edad y dudosamente de interacción (casos C y D de la figura 1).

Veamos la tabla del ANOVA de dos factores:

Variable dependiente: URICO

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	20523,989	5	4104,798	67,932	,000
Intercept	418069,431	1	418069,431	6918,765	,000
ALCOHOL	13390,636	1	13390,636	221,606	,000
EDADREC	2472,711	2	1236,355	20,461	,000
ALCOHOL * EDADREC	70,148	2	35,074	,580	,561
Error	10937,005	181	60,425		
Total	492285,000	187			
Total corregida	31460,995	186			

La evaluación de los efectos descarta la existencia de interacción significativa, al obtener un valor F = 0,58, no significativo (p=0,561). Las medias de ácido úrico son distintas dependiendo del grupo de edad y de si se es o no consumidor de alcohol, pero las diferencias de medias entre grupos de una variable se mantienen constantes para las categorías de la otra variable al no existir interacción(ver figura 4, modelo C).

Comprobamos a continuación la homogeneidad de varianzas y normalidad de los residuos:

Contraste de Levene sobre la igualdad de las varianzas

Variable dependiente: URICO

F	gl1	gl2	Significación
2,140	5	181	,063

No podemos rechazar la homogeneidad de varianzas con nivel de significación α=0,05. Aunque, como observamos en la tabla siguiente, la normalidad de los residuos debe ser rechazada.

Prueba de Kolmogorov-Smirnov para una muestra

		Residuo para URICO
N		187
Parámetros normales	Media	,0000
	Desviación típica	7,66819
Diferencias más extremas	Absoluta	,109
	Positiva	,109
	Negativa	-,090
Z de Kolmogorov-Smirnov		1,496
Sig. asintót. (bilateral)		,023

Replicando el análisis para la variable transformada SQURICO, raíz cuadrada de URICO, obtenemos:

Tabla de Anova

Variable dependiente: SQURICO

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	98,955	5	19,791	67,688	,000
Intercept	7806,255	1	7806,255	26698,321	,000
ALCOHOL	65,478	1	65,478	223,942	,000
EDADREC	11,260	2	5,630	19,256	,000
ALCOHOL * EDADREC	,430	2	,215	,735	,481
Error	52,922	181	,292		
Total	9283,000	187			
Total corregida	151,877	186			

Prueba de Kolmogorov-Smirnov para una muestra

		Residuo para SQURICO
N		187
Parámetros normales	Media	,0000
	Desviación típica	,53341
Diferencias más extremas	Absoluta	,098
	Positiva	,098
	Negativa	-,098
Z de Kolmogorov-Smirnov		1,344
Sig. asintót. (bilateral)		,054

Contraste de Levene sobre la igualdad de las varianzas error

Variable dependiente: SQURICO

F	gl1	gl2	Significación
1,476	5	181	,200

Resultados que concuerdan en la detección de efectos, no interacción y efectos de la edad y del alcohol sobre las medias de ácido úrico, y que, como se observa, permiten suponer los requerimientos de homogeneidad de varianzas (p=0,20) y normalidad (p=0,054).

El procedimiento utilizado con SPSS® para generar estos resultados ha sido el UNIVARIANTE, incluido en el MODELO LINEAL GENERAL del desplegable general del menú Analizar. La sintaxis del procedimiento es (con variable respuesta URICO):

```
UNIANOVA
  urico  BY alcohol edadrec
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /SAVE = RESID
  /PLOT = PROFILE( alcohol edadrec alcohol*edadrec )
  /PRINT = DESCRIPTIVE HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /DESIGN = alcohol edadrec alcohol*edadrec .
```

Los ajustes a la normalidad se han comprobado con la prueba de Kolmogorov-Smirnov.

La descomposición de la variabilidad en el ANOVA de dos factores

Tal como fue expuesto al inicio de este apartado, las expresiones desarrolladas en él para la descomposición de la variabilidad son válidas cuando el diseño es completo (no hay casillas vacías) y equilibrado (cumple la relación de equilibrio entre tamaños de las celdas, marginales y total). Sin embargo, el ejemplo aquí tratado es completo pero no es equilibrado, con tamaños de muestrales distintos en las celdas. ¿Puede utilizarse la descomposición de la variabilidad de la tabla del ANOVA que produce el paquete SPSS® para establecer los efectos existentes? La respuesta es que sí. Ello es debido a que el programa incorpora las rutinas para obtener los estadísticos F exactos. Para ello, en la especificación del modelo de ANOVA debe seleccionarse la descomposición de variabilidad tipo III. Las sumas de cuadrados de Tipo III tienen una gran ventaja por ser invariables respecto a la frecuencia de casillas, siempre que la forma general de estimabilidad permanezca constante. Así, este tipo de sumas de cuadrados se considera a menudo útil para un modelo no equilibrado sin casillas perdidas. El método Tipo III para la obtención de sumas de cuadrados se utiliza normalmente para cualquier modelo equilibrado o desequilibrado sin casillas vacías.

ANÁLISIS DE LA COVARIANZA

El procedimiento de análisis de la covarianza (ANCOVA) es una generalización del ANOVA y de la REGRESION LINEAL. La situación de aplicación de este procedimiento se produce cuando las variables explicativas son cualitativas y cuantitativas. En este caso el ANOVA sólo es aplicable si categorizamos las cuantitativas, perdiendo información, mientras que la regresión lineal permite introducir variables cualitativas pero en forma de variables indicador o *dummys*. Tal como se planteó en el cuadro 1, el ANCOVA permitirá comprobar si existe asociación entre la variable respuesta y un factor, controlando el posible efecto de confusión o interacción de una covariable, permitiendo obtener medidas ajustadas por la covariable de las medias de la variable respuesta en los niveles del factor.

Los datos

Sean:

Y = Variable respuesta a estudio, cuantitativa continua
X_1 = Variable explicativa, cuantitativa continua

X_2 = Variable explicativa, categórica, con I categorías

Considere como ejemplo un estudio en el que se ha medido las siguientes variables:

Y = PESO	= Peso en kilogramos	
X_1 = EDAD	= Edad en años	
X_2 = ALCOHOL	= Consumo de alcohol: 1 ‘bajo’ 2 ‘moderado/alto’	

Se dispone de una muestra con n observaciones de las variables estudiadas, $\{Y_i; X_{1i}; X_{2i}\}_{i=1}^{n}$, encontrando en la tabla 3 los datos correspondientes a 188 sujetos.

Tabla 3.- Datos de 188 observaciones de las variables peso, edad y alcohol

PESO	EDAD	ALCOHOL	PESO	EDAD	ALCOHOL	PESO	EDAD	ALCOHOL	PESO	EDAD	ALCOHOL
50	19	1	64	28	1	44	40	1	50	27	2
47	20	1	57	29	1	60	40	1	50	27	2
48	20	1	46	29	1	58	40	1	56	27	2
54	21	1	62	29	1	56	40	1	62	27	2
48	21	1	47	29	1	47	40	1	43	28	2
54	21	1	44	30	1	86	41	1	57	30	2
58	21	1	40	30	1	52	42	1	,	30	2
55	21	1	53	30	1	58	43	1	46	31	2
45	21	1	50	30	1	53	43	1	68	32	2
64	21	1	44	30	1	50	43	1	56	32	2
46	21	1	54	30	1	45	43	1	46	33	2
50	21	1	45	30	1	50	43	1	55	33	2
42	21	1	64	30	1	59	43	1	52	33	2
52	21	1	50	31	1	47	44	1	46	35	2
54	22	1	48	31	1	53	44	1	56	35	2
44	22	1	44	31	1	58	46	1	50	37	2
42	22	1	49	31	1	47	46	1	66	38	2
46	22	1	54	31	1	41	46	1	56	40	2
38	22	1	53	32	1	54	47	1	55	40	2
50	22	1	81	32	1	57	47	1	52	40	2
39	22	1	62	32	1	53	47	1	46	41	2
40	22	1	50	32	1	57	48	1	50	41	2
54	22	1	50	32	1	54	48	1	68	41	2
38	22	1	48	32	1	48	48	1	42	42	2
50	22	1	58	32	1	50	49	1	67	43	2
50	23	1	53	32	1	71	53	1	55	43	2
42	23	1	47	33	1	56	53	1	72	45	2
50	23	1	52	33	1	48	54	1	56	45	2
48	23	1	55	33	1	58	54	1	,	46	2
52	24	1	45	34	1	63	54	1	55	46	2
52	24	1	50	34	1	62	19	2	76	46	2
59	24	1	50	35	1	73	21	2	38	47	2
54	24	1	55	35	1	60	21	2	48	48	2
60	25	1	54	36	1	60	21	2	57	48	2
45	25	1	48	36	1	58	22	2	66	48	2
48	25	1	45	36	1	64	22	2	50	48	2
54	25	1	48	36	1	68	22	2	55	48	2
47	26	1	56	37	1	50	22	2	63	49	2
50	26	1	46	37	1	52	22	2	54	50	2
48	26	1	52	37	1	60	24	2	56	50	2
55	27	1	60	38	1	62	24	2	60	52	2
50	27	1	54	39	1	53	25	2	43	52	2
44	28	1	43	39	1	70	25	2	51	54	2
48	28	1	78	39	1	56	25	2	56	54	2
51	25	1	53	39	1	64	26	2	68	54	2
45	28	1	40	39	1	58	27	2	50	55	2
54	28	1	44	39	1	72	27	2	66	55	2

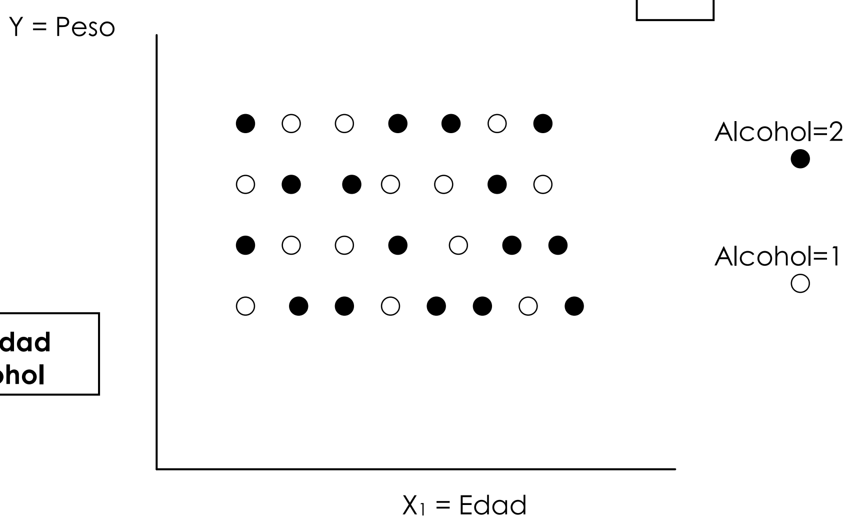
El objetivo. Modelos alternativos

Se pretende averiguar el efecto de las variables explicativas (cualitativa y cuantitativa) sobre la variable respuesta. La lectura del efecto de la variable cualitativa será a través de la existencia de diferencia de medias de la variable respuesta según sus categorías. La variable cuantitativa expresará su efecto a través de los coeficientes de un modelo de regresión

Se trata de averiguar en cuál de las siguientes situaciones nos encontramos:

i. Ninguna de las variables explicativas tiene efecto sobre la variable respuesta

A

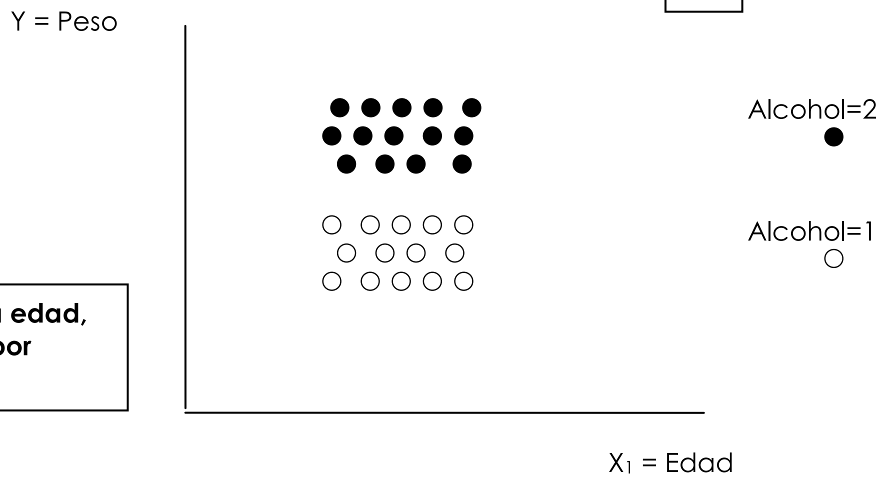


No hay efecto de la edad ni diferencias por alcohol

En esta situación, no hay modelo explicativo que dé forma a la relación entre las variables explicativas y la respuesta

ii. La variable cuantitativa no tiene efecto lineal sobre la respuesta pero ésta si que presenta diferencia en sus medias según las categorías de la cualitativa

B



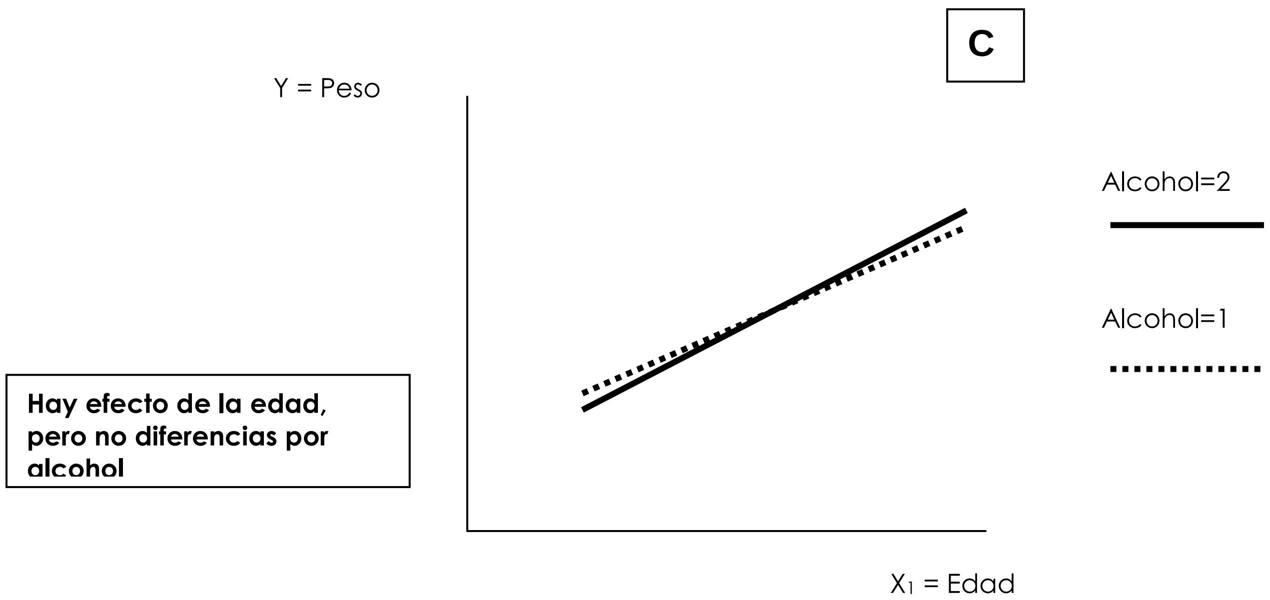
No hay efecto de la edad, pero sí diferencias por alcohol

En esta situación el modelo es un ANOVA tomando como factor la variable cualitativa. Equivalentemente, si la variable cualitativa es dicotómica o transformada en variables indicador (*dummys)*, el modelo sería un modelo de regresión

$$Y = \beta_0 + \beta_2 X_2$$

El efecto de X_2 se recoge en el parámetro β_2 y se traduce en la diferencias de medias existentes entre las categorías.

iii. La variable cuantitativa tiene efecto lineal sobre la respuesta pero no hay diferencia de medias entre las categorías de la respuesta

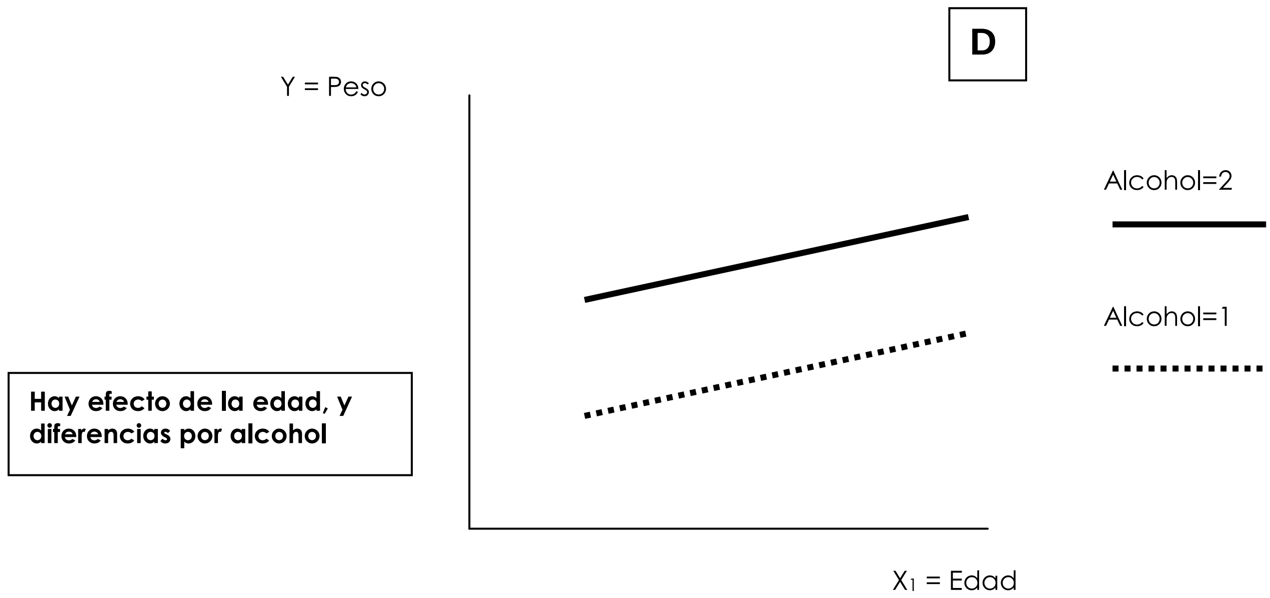


En esta situación el modelo es una única recta de regresión, común para ambas categorías de la cualitativa

$$Y = \beta_0 + \beta_1 X_1$$

El efecto de X_1 se recoge en el parámetro β_1 y se traduce en el incremento en Y por unidad de incremento en X_1 . Ajustando el correspondiente modelo de regresión lineal podemos evaluar los efectos y demás aspectos del modelo

iv. Las dos variables explicativas tienen efecto sobre la variable respuesta, pero de forma constante



En esta situación el modelo es un modelo de regresión que incorpora los efectos tanto de X_1 como de X_2

$$Y = \beta_0 + \beta_1^a X_1 + \beta_2^a X_2$$

El efecto de X_1 se recoge en el parámetro β_1^a y se traduce en el incremento en Y por unidad de incremento en X_1. El efecto de X_2 se recoge en el parámetro β_2^a y expresa la magnitud de las diferencias en las medias según las categorías. Ambos efectos están simultáneamente ajustados uno por el otro (el superíndice a representa esta situación).

Respecto a los efectos de las variables puede suceder:

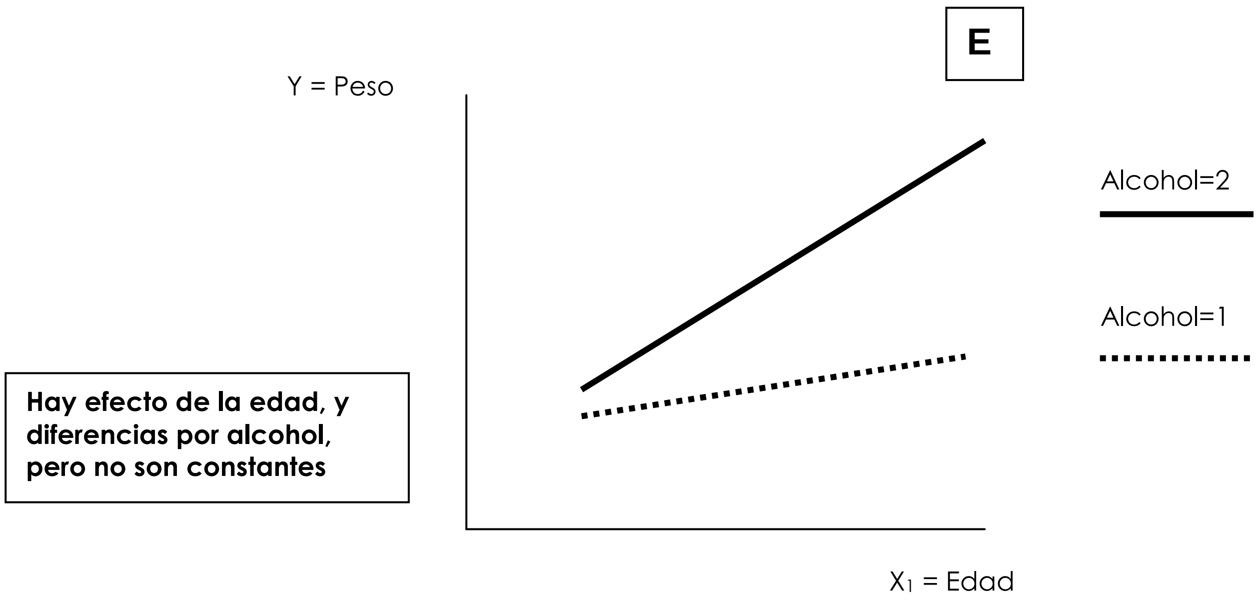
$\beta_1 = \beta_1^a$	$\Rightarrow$	El efecto de X_1 no cambia al considerar el de X_2
$\beta_1 \neq \beta_1^a$	$\Rightarrow$	Existe ***confusión***. La variable X_2 es confundiente del efecto de la variable X_1

y, de forma simétrica,

$\beta_2 = \beta_2^a$	$\Rightarrow$	El efecto de X_2 no cambia al considerar el de X_1
$\beta_2 \neq \beta_2^a$	$\Rightarrow$	Existe ***confusión***. La variable X_1 es confundiente del efecto de la variable X_2

Diremos que una variable es confundiente del efecto o relación existente entre otras dos si la magnitud de este efecto cambia al ajustarlo por ella.

v. Las dos variables tienen efecto sobre la variable respuesta pero éste no es constante



En esta situación el modelo no puede ser un único modelo de regresión. La solución sería dos modelos separados:

$$Y = \beta_0^1 + \beta_1^1 X_1$$

$$Y = \beta_0^2 + \beta_1^2 X_1$$

El efecto de X_1 no es el mismo en los dos grupos de la variable cualitativa. Simultáneamente el efecto de X_2 no es el mismo según los valores de la variable cuantitativa. Este fenómeno se conoce como ***interacción.*** Un modelo conjunto para recoger estas modificaciones de efectos podría ser:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} (X_1 \bullet X_2)$$

De acuerdo con esta situación, la diferencia en las medias de peso según consumo de alcohol no son las mismas, dependiendo de la edad de los sujetos. Por otra parte, el efecto de la edad no es el mismo en ambos grupos de consumo de alcohol.

El análisis de la covarianza consistirá en identificar la situación en la que nos encontramos y estimar los efectos sobre la variable respuesta o valores medios de ésta en diferentes situaciones.

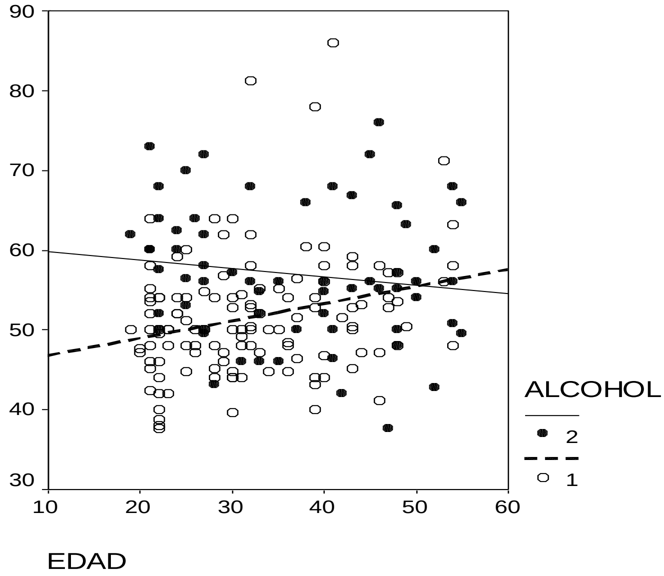
Descomposición de la variabilidad y pruebas de hipótesis

Para exponer las diferentes pruebas de hipótesis que permitirán identificar la situación utilizaremos algunos ejemplos.

Ejemplo 1. Consideremos en primer lugar los datos del ejemplo introducido al inicio de este apartado (tabla 3), con variables:

Variable respuesta Y = Peso
Variable explicativa continua X_1 = Edad
Variable explicativa categórica X_2 = Alcohol (dos categorías)

Realizaremos en primer lugar una aproximación gráfica al posible modelo para estos datos. El gráfico de dispersión conjunto tiene la forma:



¿Qué sugiere este gráfico?. Las líneas del gráfico son las rectas de regresión lineal simple que representan las tendencias de peso en función de la edad en cada uno de los grupos de alcohol. Las tendencias sugieren una posible interacción (líneas que se cruzan, gráfico tipo E en el subapartado anterior), con efectos diferentes de cada una de las variables explicativas según la categoría o valores de la otra.

Para comprobar esta situación puede ser utilizado el procedimiento UNIVARIANTE de la opción MODELO LINEAL GENERAL del desplegable Analizar del SPSS® que incorpora la opción de introducción de covariables y factores, permitiendo comprobar las hipótesis sobre los efectos existentes. La tabla siguiente muestra la salida básica del procedimiento:

Pruebas de los efectos inter-sujetos

Variable dependiente: PESO

Fuente	Suma de cuadrados tipo I	gl	Media cuadrática	F	Significación
Modelo corregido	1780,237	3	593,412	9,642	,000
Intercept	529792,249	1	529792,249	8607,837	,000
ALCOHOL	1199,527	1	1199,527	19,489	,000
EDAD	124,245	1	124,245	2,019	,157
ALCOHOL * EDAD	456,465	1	456,465	7,416	,007
Error	11201,674	182	61,548		
Total	542774,160	186			
Total corregida	12981,911	185			

En esta tabla, el programa SPSS® produce la descomposición de la variabilidad a través de la llamada suma de cuadrados de tipo I, según la cual cada uno de los efectos es añadido en el modelo a los efectos precedentes.

A través de la variabilidad explicada por cada término ALCOHOL, EDAD, INTERACCION (identificada por ALCOHOL*EDAD), puede ser resuelto el contraste:

H_0: La variable no añade variabilidad explicada a los efectos precedentes en el modelo
H_a: La variable añade variabilidad explicada a los efectos precedentes en el modelo

Así, la significación del término ALCOHOL ($p<0{,}001$) sugiere que esta variable tiene un efecto significativo. La no significación de la EDAD ($p=0{,}157$) sugiere que esta variable no añade efecto significativo al alcohol, pero la significación de la interacción ALCOHOL*EDAD ($p=0{,}007$) sugiere que ésta añade efecto significativo a los precedentes, resultando entonces que existe interacción como sugería el gráfico de dispersión. Detectada la interacción, es el efecto jerárquicamente superior y no procede comprobar la existencia de confusión u otros efectos. De la interacción se desprende que el efecto del alcohol no es constante a cualquier edad y que el efecto de la edad es distinto según el grupo de alcohol. Una forma de caracterizar el resultado puede ser obtener los modelos de regresión del peso como función de la edad en cada grupo de consumo de alcohol, obteniendo los siguientes modelos:

En ALCOHOL=1 (consumo bajo)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	44,584	2,455		18,160	,000
	EDAD	,216	,073	,259	2,962	,004

a. Variable dependiente: PESO

En ALCOHOL=2 (consumo moderado/alto)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	60,801	3,707		16,402	,000
	EDAD	-,105	,097	-,138	-1,082	,283

a. Variable dependiente: PESO

observando que la edad tiene un efecto significativo sobre el peso ($p=0{,}004$), con un incremento estimado de 0,216 kg por año, cuando el consumo de alcohol es bajo, mientras que en el grupo de consumo de alcohol moderado/alto el efecto no es significativo ($p=0{,}283$).

De forma similar, el efecto del alcohol no es el mismo según la edad. Un modelo útil para explicar este resultado puede ser el modelo de regresión que incluya un término de interacción:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} (X_1 \bullet X_2)$$

La tabla siguiente muestra el resultado al estimar este modelo:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	28,366	6,166		4,600	,000
	EDAD	,537	,177	,649	3,039	,003
	ALCOHOL	16,218	4,287	,915	3,783	,000
	ALC_EDAD	-,321	,118	-,931	-2,723	,007

a. Variable dependiente: PESO

De forma que si expresamos el modelo estimado en cada grupo de alcohol tendremos:

En ALCOHOL=1 PESO = 28,37 + 0,54 EDAD + 16,22 . 1 – 0,32 (EDAD . 1)

En ALCOHOL=2 PESO = 28,37 + 0,54 EDAD + 16,22 . 2 – 0,32 (EDAD . 2)

Y restando ambas expresiones:

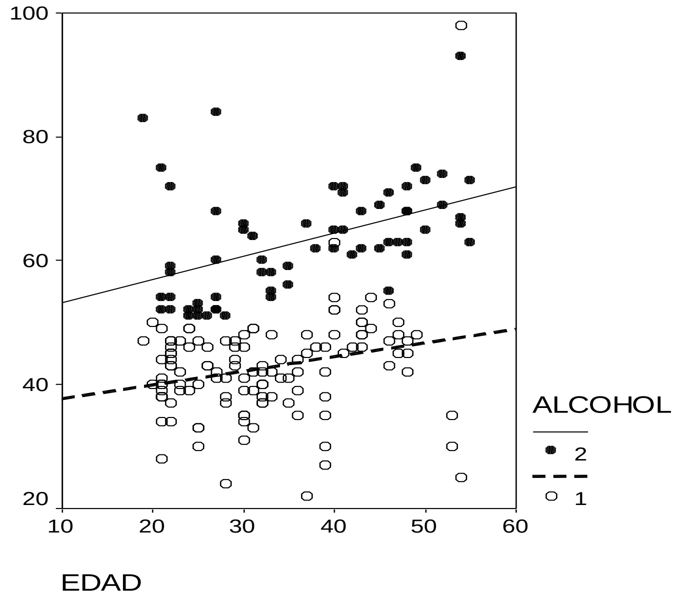
$$\Delta \text{PESO} = 16{,}22 - 0{,}32 \text{ EDAD}$$

obteniendo así que la diferencia de peso (en promedio) al comparar el grupo de consumo de alcohol moderado/alto frente al bajo depende de la edad. Según este resultado, el efecto del alcohol disminuye con la edad, es decir, es mayor en los más jóvenes y es menor en los más mayores. Por ejemplo, con EDAD = 30 obtenemos ΔPESO = 6,62, mientras que con EDAD = 50 obtenemos ΔPESO = 0,22.

<u>Ejemplo 2</u>. Como segundo ejemplo consideremos las variables:

Variable respuesta Y = Acido úrico
Variable explicativa continua X_1 = Edad
Variable explicativa categórica X_2 = Alcohol (dos categorías

con datos de ácido úrico descritos en la tabla 1. La aproximación gráfica produce el siguiente diagrama de dispersión



Las tendencias de las líneas de regresión nos sitúan según el subapartado anterior en un posible modelo D (efectos de ambas variables) o E (interacción). Veamos la tabla de descomposición de variabilidades:

Pruebas de los efectos inter-sujetos

Variable dependiente: URICO

Fuente	Suma de cuadrados tipo I	gl	Media cuadrática	F	Significación
Modelo corregido	19473,155	3	6491,052	99,089	,000
Intercept	460824,005	1	460824,005	7034,695	,000
ALCOHOL	17853,570	1	17853,570	272,543	,000
EDAD	1520,449	1	1520,449	23,210	,000
ALCOHOL * EDAD	99,136	1	99,136	1,513	,220
Error	11987,840	183	65,507		
Total	492285,000	187			
Total corregida	31460,995	186			

Podemos observar que el efecto de interacción (ALCOHOL*EDAD) no aporta variabilidad significativa a los precedentes, mientras que éstos si son significativos. Descartada la interacción, construimos un modelo de regresión para obtener los efectos ajustados de ambas variables.

Veamos en primer lugar los modelos de regresión simple en función de cada una de las variables:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			Intervalo de confianza para B al 95%	
		B	Error típ.	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	21,999	1,882		11,691	,000	18,287	25,712
	ALCOHOL	20,594	1,322	,753	15,580	,000	17,986	23,202

a. Variable dependiente: URICO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			Intervalo de confianza para B al 95%	
		B	Error típ.	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	33,434	3,113		10,740	,000	27,292	39,576
	EDAD	,481	,089	,371	5,431	,000	,306	,655

a. Variable dependiente: URICO

Ambas variables tienen, solas, efecto significativo sobre el ácido úrico. Veamos el modelo de regresión múltiple

$$Y = \beta_0 + \beta_1^a X_1 + \beta_2^a X_2$$

para ajustar los de cada variable por la otra:

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	19374,019	2	9687,009	147,465	,000[a]
	Residual	12086,976	184	65,690		
	Total	31460,995	186			

a. Variables predictoras: (Constante), EDAD, ALCOHOL

b. Variable dependiente: URICO

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Coeficientes no estandarizados Error típ.	Coeficientes estandarizados Beta	t	Sig.	Intervalo de confianza para B al 95% Límite inferior	Intervalo de confianza para B al 95% Límite superior
1	(Constante)	13,883	2,451		5,663	,000	9,047	18,719
	ALCOHOL	19,324	1,277	,707	15,135	,000	16,805	21,843
	EDAD	,291	,061	,225	4,811	,000	,172	,411

a. Variable dependiente: URICO

en el que observamos el ajuste significativo del modelo (tabla de ANOVA) y los efectos significativos de ambas variables (tabla de coeficientes).

Para discutir el efecto de confusión deberemos comparar el efecto de cada variable sóla con su efecto ajustado por la otra variable, es decir, comparar los coeficientes de cada variable en el modelo simple con los coeficientes en el modelo múltiple:

Variable	Coeficiente β simple	Coeficiente β^a ajustado
ALCOHOL	20,59 $IC_{0,95}$=[17,99 ; 23,20]	19,32 $IC_{0,95}$=[16,81 ; 21,84]
EDAD	0,48 $IC_{0,95}$=[0,31 ; 0,66]	0,29 $IC_{0,95}$=[0,17 ; 0,41]

observando una disminución de los efectos de ambas variables al ajustar por la otra aunque si observamos los intervalos de confianza (como guía aproximada para establecer el nivel de confusión) vemos que se solapan.

El cálculo de las medias ajustadas

Cuando no hay interacción (ejemplo 2 del subapartado anterior) podemos estar interesados en estimar los valores de las medias de la variable respuesta en los niveles del factor ajustadas por la covariable. Nótese que no nos referimos al efecto del factor (diferencia de medias) ajustado por la covariable, pues éste se encuentra en el parámetro β^a del factor en el modelo de regresión múltiple. Se trata de responder la pregunta ¿cuánto valdrían las medias de ácido úrico según consumo de alcohol si los individuos de los grupos de consumo de alcohol tuvieran la misma edad media? Esta pregunta tiene sentido cuando las edades medias de los grupos no son iguales y cuando la edad tiene un efecto significativo. En el ejemplo 2 analizado hemos encontrado un efecto significativo de la edad, y, además, la edad media presenta diferencias entre los niveles de consumo de alcohol:

EDAD

ALCOHOL	Media	N	Desv. típ.
1	32,22	123	9,144
2	36,58	64	11,065
Total	33,71	187	10,031

Podemos estimar las medias de ácido úrico ajustadas por edad a través del modelo de regresión como:

$$\text{Media ajustada por edad}_{/\text{alcohol}} = \beta_0 + \beta_1^a \,\text{EDAD MEDIA} + \beta_2^a \,\text{ALCOHOL}$$

A partir del modelo obtenido en el subapartado anterior, con SPSS® obtenemos los siguientes resultados para las medias ajustadas en cada grupo de alcohol:

Estimaciones

Variable dependiente: URICO

ALCOHOL	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
1	43,028[a]	,736	41,575	44,481
2	62,352[a]	1,028	60,324	64,380

a. Evaluado respecto a cómo aparecen las covariables en el modelo: EDAD =33,71.

resultados que representan las medias de ácido úrico estimadas para cada grupo de alcohol (con sus correspondientes intervalos de confianza) si la edad media de los sujetos fuera la misma en cada grupo (33,71 años).

Requerimientos. Aplicación de procedimientos con SPSS®

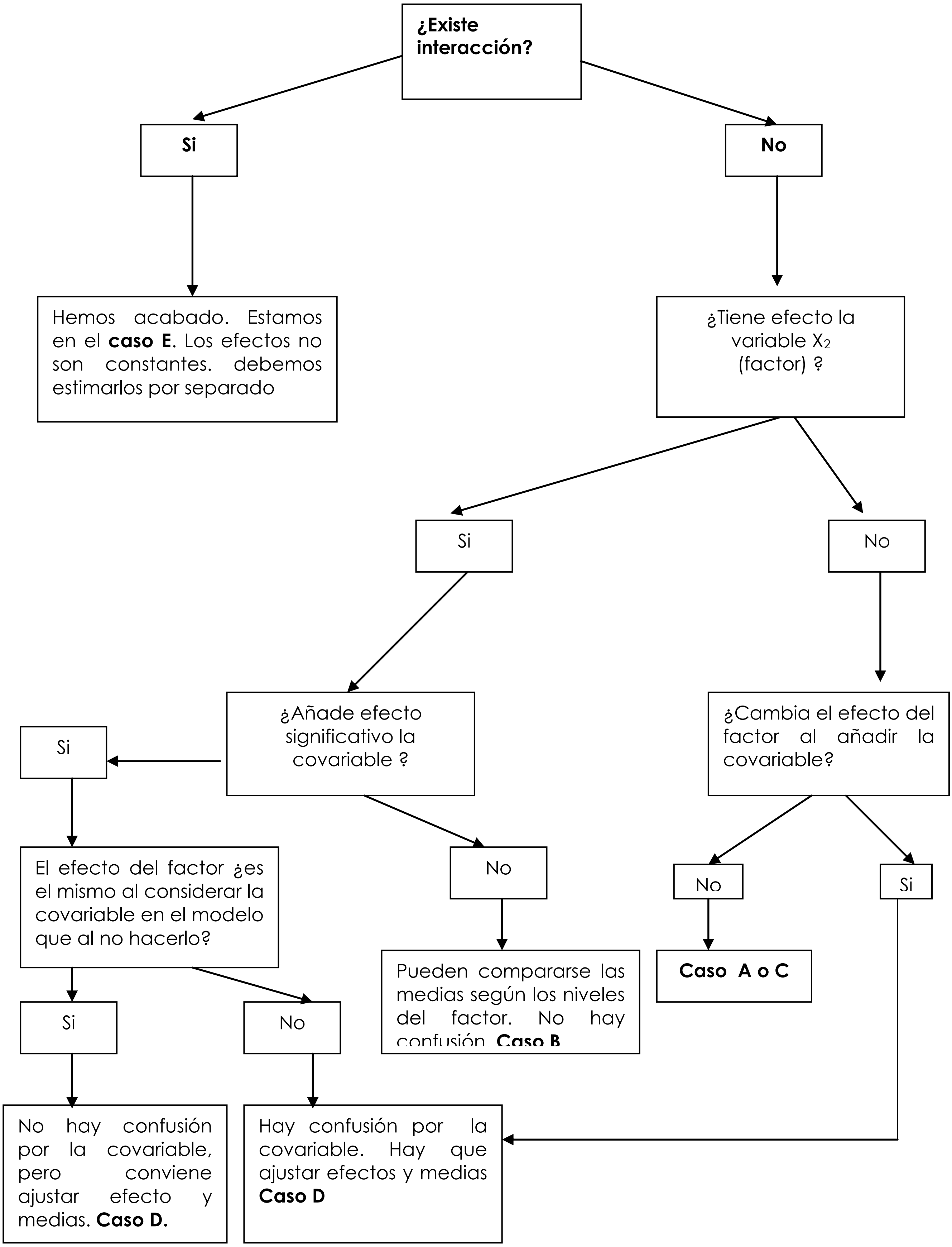
Los requerimientos del ANCOVA son los mismos descritos para el ANOVA, a saber normalidad, homogeneidad e independencia. Su comprobación se puede realizar de forma similar al ANOVA, pues el procedimiento utilizado con SPSS® es el mismo UNIVARIANTE del MODELO LINEAL GENERAL incluido en el desplegable Analizar. La sintaxis del procedimiento utilizado para resolver los contrastes ha sido:

```
UNIANOVA
  urico  BY alcohol  WITH edad
  /METHOD = SSTYPE(1)
  /INTERCEPT = INCLUDE
  /SAVE = RESID
  /PLOT = PROFILE( alcohol )
  /EMMEANS = TABLES(alcohol) WITH(edad=MEAN)
  /PRINT = DESCRIPTIVE ETASQ HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /DESIGN = alcohol edad alcohol*edad .
```

Una estrategia de análisis

Cuando el papel de la variable cualitativa es el de factor y la variable cuantitativa actúa de covariable (ver apartado 'Introducción. Conceptos generales'), podemos actuar de diversas formas para detectar los efectos existentes y producir las estimaciones correspondientes. En la figura 5 podemos observar una posible estrategia de análisis a seguir en esta situación. Los modelos resultantes se identifican según los posibles modelos de ANCOVA (ver subapartado 'El objetivo. Modelos alternativos')

Figura 5.- Estrategia de análisis en ANCOVA con un factor cualitativo y una covariable


¿Existe interacción?
Si
No
Hemos acabado. Estamos en el caso E. Los efectos no son constantes. debemos estimarlos por separado
¿Tiene efecto la variable X2 (factor) ?
Si
No
¿Añade efecto significativo la covariable ?
¿Cambia el efecto del factor al añadir la covariable?
Si
El efecto del factor ¿es el mismo al considerar la covariable en el modelo que al no hacerlo?
No
Pueden compararse las medias según los niveles del factor. No hay confusión. Caso B
No
Si
Caso A o C
Si
No
No hay confusión por la covariable, pero conviene ajustar efecto y medias. Caso D.
Hay confusión por la covariable. Hay que ajustar efectos y medias Caso D

ANÁLISIS DE LA VARIANZA DE MEDIDAS REPETIDAS

Los procedimientos de ANOVA de uno o más factores requieren que las observaciones hayan sido generadas de forma independiente. Cuando el diseño es de tipo apareado, o no independiente, tales procedimientos no son aplicables. Un tipo de diseño apareado se produce cuando las observaciones de nuestros datos corresponden a los mismos sujetos pero observados en diferentes momentos. Diremos entonces que nos encontramos en una situación de medidas repetidas (para cada sujeto)

Los datos

Sean:

n = número de sujetos a estudio
Y = Variable respuesta a estudio, cuantitativa continua
Y_1 = Observaciones de la variable Y en el instante 1
Y_2 = Observaciones de la variable Y en el instante 2
.
.
Y_k = Observaciones de la variable Y en el instante K

Disponiendo de una muestra de n observaciones repetidas k veces de la variable Y, distribuidas según la siguiente estructura:

	Momento de observación			
Sujeto	1	2	...	k
1	Y_{11}	Y_{12}		Y_{1k}
2	Y_{21}	Y_{22}		Y_{2k}
3	Y_{31}	Y_{32}		Y_{3k}
.	.	.		.
.	.	.		.
.	.	.		.
n	Y_{n1}	Y_{n2}		Y_{nk}

donde $\{Y_{ij}\}_{j=1}^{k}$ representa el conjunto de k medidas repetidas realizadas sobre el sujeto i, y μ_j es la media poblacional de la variable Y en el momento j.

Consideremos como ejemplo un estudio realizado sobre 12 sujetos, a los que se les ha realizado determinación de su nivel de ácido úrico tres veces:

- Al inicio de un tratamiento consistente en una dieta (momento 1)
- Tras 1 mes de tratamiento (momento 2)
- Tras 3 meses de tratamiento (momento 3)

Las observaciones obtenidas son las siguientes:

	Momento de observación		
Sujeto	1	2	3
1	6,5	5,9	7,2
2	5,0	5,5	6,2
3	6,4	6,4	7,0
4	7,2	7,8	6,4
5	9,2	7,1	6,5
6	8,0	7,0	7,0
7	7,5	5,0	5,1
8	9,2	6,5	5,1
9	6,9	6,9	5,5
10	7,8	6,0	5,9
11	7,0	4,0	4,2
12	6,5	6,3	6,2
$\overline{X}$	7,3	6,2	6,0

El objetivo. Transformaciones sobre las variables originales

A partir del ejemplo planteado en el subapartado anterior, sean

URICO1 = Variable que recoge las observaciones en el momento 1
URICO2 = Variable que recoge las observaciones en el momento 2
URICO3 = Variable que recoge las observaciones en el momento 3

Se pretende resolver la prueba de hipótesis:

H_0 : No hay diferencia en las medias de ácido úrico $\mu_1 = \mu_2 = \mu_3$

H_a : Las medias de ácido úrico difieren entre momentos, $\exists\ \mu_j$ diferente a los demás

Para comprobar las hipótesis planteadas las variables originales deben ser transformadas, es decir, en lugar de analizar esta variables se analizará combinaciones lineales de sus diferencias que se denominan **contrastes.** Los contrastes dan lugar a variables nuevas que son **independientes** (ortogonales) y tienen la propiedad de ser **normalizados** (la suma de los cuadrados de sus coeficientes es 1). Tales contrastes se denominan **ortonormalizados.** El número de contrastes ortonormalizados es igual al de momentos a estudio (o niveles del factor momento) menos 1. Para nuestros datos se pueden generar 2 contrastes ortonormalizados.

Tomando como referencia el ejemplo descrito, presentamos a continuación algunos tipos de transformaciones de las variables originales en contrastes, concretamente las que incluye el programa SPSS®. Sea cual sea la transformación elegida, siempre se calcula adicionalmente a los contrastes propios de esa transformación el llamado contraste PROMEDIO, transformación global que sirve para comprobar si la media global μ de los datos es 0 o no:

$$\text{CONTRASTE PROMEDIO} = (\text{URICO1} + \text{URICO2} + \text{URICO3}) / \sqrt{3}$$

i. Contrastes 'polinómicos'

Se crean las siguientes variables, cada una de ellas recogiendo un efecto:

V1 = Promedio = (URICO1 + URICO2 + URICO3) / $\sqrt{3}$

V2 = Lineal = (URICO3 – URICO1) / $\sqrt{2}$
V3 = Cuadrático = (URICO3 – 2. URICO2 + URICO1) / $\sqrt{6}$

Esta variables serán utilizadas para las pruebas de hipótesis que se presentarán después. Los efectos que permitirán detectar son:

V1 = Comprobación sobre si la media global es 0 o no
V2 = Comprobación de si las diferencias de medias siguen una tendencia lineal
V3 = Comprobación de si las diferencias de medias siguen una tendencia curvilínea (baja-sube-baja o sube-baja-sube)

ii. Contrastes tipo 'diferencia'

Las variables creadas son:

V1 = Promedio
V2 = URICO2-URICO1
V3 = URICO3 – $\frac{1}{2}$ (URICO1 + URICO2) = URICO3 – 0,5 URICO1 – 0,5 URICO2

Las variables V2 y V3 se usan para detectar si las diferencias se producen entre URICO2 y URICO1 y/o entre URICO3 y la media de los precedentes

iii. Contrastes tipo 'desviación'

Las variables creadas son:

V1 = Promedio
V2 = URICO1- $\frac{1}{3}$ (URICO1+URICO2+URICO3)
V3 = URICO2 – $\frac{1}{3}$ (URICO1 + URICO2 + URICO3)

Las variables V2 y V3 se usan para detectar si las diferencias se producen entre URICO2 y la media y/o URICO1 y la media de los tres

iv. Contrastes tipo 'Helmert'

La construcción es idéntica a la de tipo 'diferencia' pero en sentido opuesto a crear las diferencias

V1 = Promedio
V2 = URICO1– $\frac{1}{2}$ (URICO2 + URICO3)
V3 = URICO2 – URICO3

La utilización es semejante a la de tipo diferencia

v. Contrastes tipo 'repetido'

Las variables creadas son:

V1 = Promedio V2 = URICO2 – URICO1 V3 = URICO3 – URICO2

Las variables V2 y V3 permitirán detectar si las diferencias entre medias se producen entre URICO2 y URICO1 y/o URICO3 y URICO2

vi. Contrastes 'simples'

Las variables creadas son:

V1 = Promedio V2 = URICO2 – Nivel de referencia = URICO2 – URICO1 V3 = URICO3 – Nivel de referencia = URICO3 – URICO1

Elegida una categoría de referencia, se comprueba si las diferencias se atribuyen a cada una de las diferencias de las otras con esta categoría de referencia

Como se observa, la elección de uno u otro contraste permite comprobar la existencia de un tipo u otro de diferencias en la variable respuesta entre los momentos a estudio

Requerimientos del ANOVA de medidas repetidas

El ANOVA de medidas repetidas es menos robusto (más influenciable) que el ANOVA por falta de cumplimiento de los requerimientos básicos, que a los ya descritos de normalidad y homogeneidad de varianzas añade el de covarianzas nulas:

i. Normalidad

ii. Homogeneidad de varianzas

iii. Covarianzas nulas para las variables transformadas como contrastes

Aplicación sobre un ejemplo

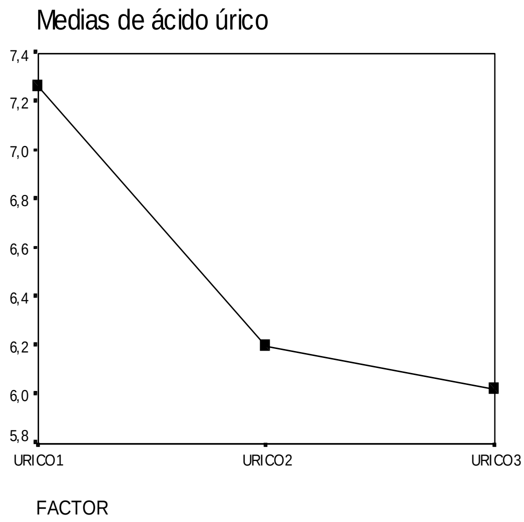
Utilizaremos el ejemplo descrito en la introducción. Denominaremos por

URICO1 = Obervaciones de úrico en el momento 1
URICO2 = Obervaciones de úrico en el momento 2
URICO3 = Obervaciones de úrico en el momento 3

Veamos en primer lugar los resultados descriptivos y gráficos:

Estadísticos descriptivos

	Media	Desv. típ.	N
URICO1	7,267	1,192	12
URICO2	6,200	1,024	12
URICO3	6,025	,907	12



Observamos una disminución de las medias muestrales desde el inicio al momento final.

La solución incluye siempre un contraste basado en la variable transformada V1 que es la misma con independencia del método elegido para la generación de las variables contraste:

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1
Variable transformada: Promedio

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	1519,700	1	1519,700	949,648	,000
Error	17,603	11	1,600		

El efecto recogido en la variable Promedio permite contrastar la hipótesis nula de que la media global (todas las observaciones juntas) es 0 o no. En este caso el resultado conduce a su rechazo ($F=949,648$; $p<0,001$)(nótese que $n=12$). En realidad este contraste no tenía iteres en esta situación.

La existencia de diferencias en las medias de ácido úrico según el momento puede ser contrastada a través de dos procedimientos: procedimiento multivariante y procedimiento univariante. Cuando ambos procedimientos coinciden la elección de uno u otro no es muy relevante. Pero cuando conducen a resultados diferentes hay que tener precaución en la elección de uno u otro. Los estadísticos del procedimiento multivariante suponen que las medidas sobre los sujetos son una muestra aleatoria de una normal multivariante y no requiere suposiciones sobre la estructura de la matriz de varianzas-covarianzas. Por otra parte, los estadísticos univariantes requiere que la matriz de varianzas-covarianzas sea la matriz identidad. Si esta suposición es aceptable, el procedimiento univariante es preferible al multivariante, especialmente con tamaños de muestra pequeños. Hay que decir que los resultados de los estadísticos por uno u otro procedimiento no dependen del tipo de variables de contraste seleccionado. Veamos a continuación los resultados de ambos procedimientos:

Estadísticos multivariantes:

Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación
URI	Traza de Pillai	,431	3,791[a]	2,000	10,000	,060
	Lambda de Wilks	,569	3,791[a]	2,000	10,000	,060
	Traza de Hotelling	,758	3,791[a]	2,000	10,000	,060
	Raíz mayor de Roy	,758	3,791[a]	2,000	10,000	,060

a. Estadístico exacto

Todos los estadísticos presentan una p=0,06. Con un criterio basado en α = 0,05, no deberíamos rechazar la hipótesis nula de igualdad en las medias de ácido úrico. Los estadísticos suponen normalidad multivariante de las medidas de los sujetos.

Estadísticos univariantes:

La prueba de esfericidad de Mauchly contrasta la hipótesis nula de que las varianzas son constantes y las covarianzas entre las variables transformadas en las variables contraste son incorreladas (covarianza 0). En este caso no puede rechazarse esta hipótesis, asumiendo que se cumple.

Prueba de esfericidad de Mauchly

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Significación	Epsilon: Greenhouse-Geisser	Epsilon: Huynh-Feldt	Epsilon: Límite-inferior
URI	.603	5.059	2	.080	.716	.793	.500

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

La tabla siguiente muestra estadísticos univariantes. A parte del estadístico F, se calculan estadísticos corregidos. Las correcciones se realizan sobre los grados de libertad, y cada una de ellas, Greenhouse-Geisser, Huynh-Feldt, y Límite inferior utilizan criterios más o menos conservadores[3]. El Límite inferior es el más conservador, le sigue Greehouse-Geisser y por último Huynh-Feldt.

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
URI	Esfericidad asumida	10.841	2	5.420	6.409	.006
	Greenhouse-Geisser	10.841	1.432	7.572	6.409	.015
	Huynh-Feldt	10.841	1.586	6.834	6.409	.012
	Límite-inferior	10.841	1.000	10.841	6.409	.028
Error(URI)	Esfericidad asumida	18.606	22	.846		
	Greenhouse-Geisser	18.606	15.747	1.182		
	Huynh-Feldt	18.606	17.450	1.066		
	Límite-inferior	18.606	11.000	1.691		

Todos los contrastes conducen a rechazar la hipótesis nula de igualdad de medias.

[3] En estadística se dice que un criterio es conservador si tiende a aceptar hipótesis nulas con mayor probabilidad de la que establece el nivel de significación definido

A partir de aquí, la interpretación de los resultados depende del método elegido en la generación de las variables contraste:

Polinómico

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	FACTOR	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
FACTOR	Lineal	9,250	1	9,250	7,069	,022
	Cuadrático	1,590	1	1,590	4,153	,066
Error(FACTOR)	Lineal	14,395	11	1,309		
	Cuadrático	4,212	11	,383		

Si decidimos que hay diferencias significativas en las medias, éstas siguen una tendencia lineal de forma significativa (p=0,022) pero no cuadrática (p=0,066). Lo cierto es que la tendencia en los datos es más cuadrática que lineal, pero no se encuentra significación suficiente para esta última (n=12)

Diferencias

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	FACTOR	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
FACTOR	Nivel 2 - Nivel 1	13,653	1	13,653	8,098	,016
	Nivel 3 - Anterior	6,021	1	6,021	4,731	,052
Error(FACTOR)	Nivel 2 - Nivel 1	18,547	11	1,686		
	Nivel 3 - Anterior	13,999	11	1,273		

Hay diferencias significativas entre el momento 2 y el 1, y casi (p=0,052) entre el momento 3 y el promedio de los anteriores

Helmert

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	FACTOR	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
FACTOR	Nivel 1 - Anterior	15,985	1	15,985	8,161	,016
	Nivel 2 - Nivel 3	,368	1	,368	,477	,504
Error(FACTOR)	Nivel 1 - Anterior	21,547	11	1,959		
	Nivel 2 - Nivel 3	8,483	11	,771		

Hay diferencias significativas entre el momento 1 y el promedio de los demás, y pero no las hay entre el momento 2 y 3 (p=0,504)

Repetido

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	FACTOR	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
FACTOR	Nivel 1 - Nivel 2	13,653	1	13,653	8,098	,016
	Nivel 2 - Nivel 3	,368	1	,368	,477	,504
Error(FACTOR)	Nivel 1 - Nivel 2	18,547	11	1,686		
	Nivel 2 - Nivel 3	8,483	11	,771		

Hay diferencias significativas entre el momento 1 y el 2 pero no entre el 2 y el 3

Simple

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	FACTOR	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
FACTOR	Nivel 1 - Nivel 3	18,501	1	18,501	7,069	,022
	Nivel 2 - Nivel 3	,368	1	,368	,477	,504
Error(FACTOR)	Nivel 1 - Nivel 3	28,789	11	2,617		
	Nivel 2 - Nivel 3	8,483	11	,771		

Tomando como referencia el momento 3, hay diferencia entre el momento 1 y el 3 pero no entre el 2 y el 3

Desviación

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	FACTOR	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
FACTOR	Nivel 1 - Media	7,105	1	7,105	8,161	,016
	Nivel 2 - Media	1,060	1	1,060	4,153	,066
Error(FACTOR)	Nivel 1 - Media	9,577	11	,871		
	Nivel 2 - Media	2,808	11	,255		

Hay diferencia entre el momento 1 y la media de los tres momentos pero no entre el momento 2 y la media (casi pues p=0,066)

Aplicación de procedimientos con SPSS®

El procedimiento utilizado para producir los resultados del ejemplo es el MEDIDAS REPETIDAS del procedimiento MODELO LINEAL GENERAL del desplegable Analizar. La sintaxis básica es:

```
GLM
  urico1 urico2 urico3
  /WSFACTOR = uri 3 Polynomial
  /METHOD = SSTYPE(3)
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = uri .
```

BIBLIOGRAFIA DE REFERENCIA

1. ANDERSON S, AUQUIER A, HAUCK W, OAKES D, VANDAELE W, WEISBERG H. Statistical Methods for comparative studies. Techniques for bias reduction. New york: John Wiley & Sons. 1980.

   Libro dedicado a introducir conceptos y procedimientos en estudios comparativos. Los primeros capítulos revisan conceptos relacionados con la medición de los efectos, los sesgos, la confusión y la interacción o la asignación aleatoria y el apareamiento. El capítulo 8 está dedicado al análisis de la covarianza.

2. FISHER LI D, VAN BELLE G. BIOSTATISTICS: A methodology for the health sciences. New York: John Wiley & Sons. 1993

   Este libro es un excelente compendio de procedimientos de estadística, con aplicación sobre ejemplos y situaciones biomédicas y sanitarias. Su capítulo 10 está dedicado al ANOVA (incluye de una vía, de dos, de medidas repetidas y ANCOVA), y establece con claridad los conceptos básicos de estos procedimientos. Dedica su capítulo 8 a técnicas no paramétricas, incluidas las alternativas al ANOVA.

3. KLEINBAUM DG, KUPPER LL, MULLER KE. Applied Regression Analysis and other multivariable methods.Boston: PWS-KENT Publishing Company. 1988

   Los capítulos 15, 18, 19 y 20 están dedicados a ANOVA y ANCOVA. Su capítulo 11 discute los conceptos de interacción y *confounding*

4. MARTIN A, LUNA DEL CASTILLO JD. BIOESTADISTICA para las Ciencias de la Salud. Madrid: Ed. Norma. 1989

   Texto general de bioestadística. Su capítulo 11 está dedicado al ANOVA, presentando la técnica y aplicándolo sobre ejemplos biomédicos. Incluye técnicas no paramétricas. El capítulo 12, dedicado a la regresión lineal, incluye el análisis de la covarianza como problema de comparación de rectas de regresión.

5. NETER J, WASSERMAN W, WHITMORE GA. Applied statistics. Boston: Allyn and Bacon. 1993.

   Libro muy aplicado. Aunque sus ejemplos son muy variados y de todos los campos del saber, su lenguaje es muy actual y es muy didáctico. Está muy bien presentada la regresión lineal a la que dedica varios capítulos. El capítulo 21 está dedicado al ANOVA

6. NORUSIS MJ. SPSS Advanced Statistics.

   Es el manual de referencia del módulo de estadísticas avanzadas del SPSS. Contiene aplicaciones de los procedimientos de ANOVA, ANCOVA y ANOVA de medidas repetidas, así como la descripción de los comandos para procesarlas.

7. PEÑA D. Estadística, Modelos y Métodos: 2. Modelos lineales y series temporales. Madrid: Alianza Editorial. 1987

   Texto general de estadística. Incluye varios capítulos dedicados al ANOVA. Su lenguaje es más matemático. No incluye aplicaciones sanitarias.